Towards principled learner selection

Andrew Rattray, BSc.

Thesis submitted to the University of Nottingham
for the degree of Master of Research

January 2013

Abstract

Long learner evaluation times are no longer exceptional and often there is insufficient time to exhaustively test all candidate options. When deciding which learners to use, practitioners must rely on ad hoc testing and luck to identify the most accurate one. Given the importance of classification in decision making, this is unsatisfactory. Progress towards a principled approach requires accurate predictions of learner accuracy and evaluation time and this study examines the potential of traditional meta-learning approaches, with their emphasis on indirect explanatory variables, to deliver the required solutions.

Here, 57 different indirect dataset characteristics, including those related to geometrical complexity, are used as explanatory variables, alongside sample-estimates, in building regression models of accuracy and time.

The evidence presented firmly suggests that these indirect variables lack both the required predictive power and the time efficiency required for the development of practically useful models, and points instead towards basing the prediction of learner accuracy solely on sample-based models. The attempt at modelling learner evaluation time reveals some of the difficulties that this tough challenge presents.

Acknowledgements

I wish to thank my supervisor, Dr Jaume Bacardit, for his input and guidance but mostly I want to thank Victoria and my Mum for their support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and motivation

Machine learning algorithms can only function because of their inductive biases Mitchell (1997) which differ between learners, leading to potential changes in relative performance from one dataset to another. The work of Wolpert (1992) and Schaffer (1994) provides the basis for believing that there is no hierarchy of learners in respect of classification accuracy. Consequently, when presented with a new dataset we need to evaluate all available learners in order to be assured of finding the most accurate.

Datasets large enough to have evaluation times of the order of hours per learning option are no longer exceptional (see Fig 1.1). Increasingly, evaluating more than a handful of the potentially hundreds of candidate learning options is not viable and deciding which learners to exclude is based on a variety of ad-hoc approaches. It is arguable that, as datasets get larger, evaluation times longer and a smaller proportion of options are evaluated, the most accurate learning options are increasingly not being identified. Time efficient, reliable predictions of learner accuracy and evaluation time are the foundation upon which a more rational, principled, selection process can be developed. Given the central role played by classification in decision making in many important areas of modern life, seeking to improve learner selection is certainly a goal worth pursuing.

Substantial research effort has been expended in developing such predictive processes using explanatory variables whose values are independent of direct learner interaction. Whether these indirect, 'meta-learning', approaches have the potential to contribute to practical learner selection is debateable, yet related studies continue to appear.

1.2 Aims and scope

This research explores the viability of typical meta-learning approaches, seeking to determine whether time expended in calculating indirect explanatory variables would be better spent evaluating learners directly.

It is also investigates whether using sample estimates of different learners in the same model can improve upon the predictive accuracy of using them individually, outside of a model. This could offer an alternative approach to meta-learning for predicting learner accuracy.

Predicting evaluation time is as important to learner selection as predicting accuracy and yet there are significantly fewer published studies. This work will present new experimental results demonstrating the difficulty of producing reliable estimates of evaluation time, regardless of the explanatory variables used.

1.3 Contributions of this thesis

1. A first comparative study of the impact of the different types of proposed meta-learning attribute on the predictive accuracy of error-rate models

2. A novel approach to analysing the efficiency with which meta-learning approaches use potential evaluation time

3. A novel approach to assessing the practical usefulness of accuracy and evaluation time predictive processes

4. A demonstration that regression modelling can exploit synergy between sample estimates for different learners

1.4 Structure of the thesis

There are five chapters. Chapter 2 explains the learner selection problem more fully before providing a review of the literature, highlighting the distinction between sample and meta-learning attempts to solve it. The knowledge gap identified in the literature drives the experimental aims and design which are presented in Chapter 3. Chapter 4 will present and discuss the experimental results. The thesis is concluded with some discussion in Chapter 5 about what has been learned and how this work can be built upon.

1.4.1 Vocabulary and abbreviations

To maintain flow, I will avoid repeatedly making the distinction between learning algorithm and induced classifier and refer to both as 'learners'. Occasionally, the term learner should be interpreted more broadly as 'learning option', encompassing the possible combinations of learner, pre-processing technique and parameter settings – it should be clear from the context when this extension should be made.

Statistical parlance ('explanatory variable', 'regression model' or plain 'model') will be used when discussing predictive processes for accuracy and evaluation time, rather than machine learning parlance ('meta-attribute', 'meta-learner'). However, the terms 'meta-learning' and 'meta-dataset' will be used occasionally when doing so saves words.

Generally, 'accuracy' will be used instead of 'error-rate', to avoid possible confusion with the predictive error-rates at the meta-level. Results, however, are presented as error-rates.

The following abbreviations will be used:

- CV (Cross-validation)
- MAE (Mean absolute error)
- SSI (Structural, statistical and information-theory based)

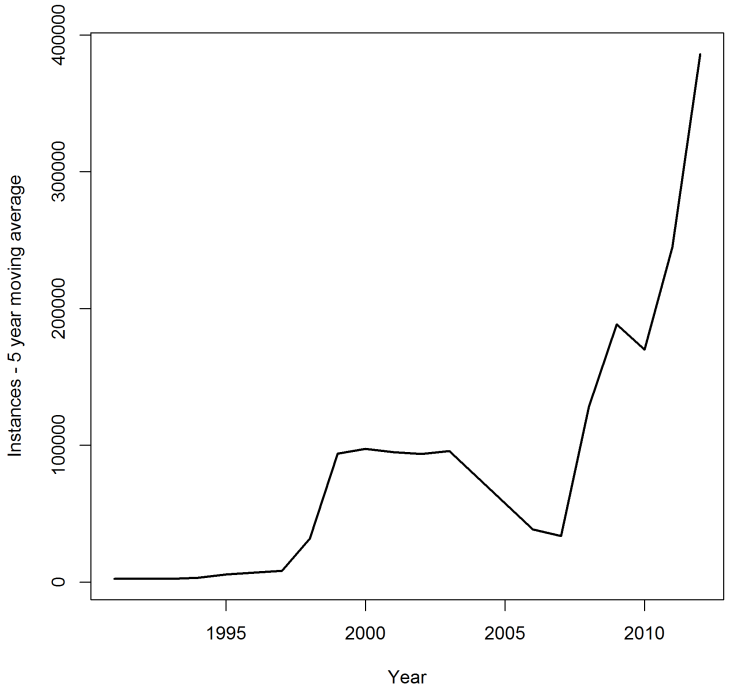

Figure 1.1: Dataset sizes and evaluation times are growing - average UCI repository dataset size by year

Chapter 2

Background and related work

In this chapter the thesis is set in context. We start by considering, in greater detail, the problem which motivates this research. The related work is divided into three categories. Sample-based and meta-learning approaches to learner accuracy are reviewed in Section 2, whilst the prediction of evaluation time is considered in Section 3. Section 4 highlights the knowledge gaps that this work aims to address.

2.1 The challenge of practical learner selection

As datasets increase in size, with corresponding increases in evaluation times, data-mining practitioners are more frequently faced with having to decide which learners (and supporting techniques) to try on a dataset and which to leave on the shelf.

When selecting appropriate learners to use on a dataset, the following criteria influence decision-making:

1. Classification accuracy
2. Evaluation (training and testing) time
3. Interpretability of classifier output
4. Experience of previous use
5. Domain pedigree of the learners
6. Computational resource limitations

Accuracy must always be a high priority but the most accurate may not be selected if others with similar accuracy perform better against other criteria.

Trade-offs between accuracy and evaluation time are central to learner selection; herein interest is confined to these two criteria.

There is a distinction between fully-informed trade-offs and those based on uncertain beliefs. In the former, the practitioner has evaluated all learners, possesses accuracy and evaluation times for each, and (assuming a large enough dataset) is able to generalise that these relative performances will hold for all new data from that domain. Any of the other criteria can then be sensibly weighed against loss of accuracy. When there is insufficient time to evaluate all learners, the most accurate may never be known and poor trade-offs become more likely. The nature of trade-offs demands numerical time and accuracy information — knowing that one learner is more accurate than another is not enough, we need to know by how much. It is upon this logic that the contention is made that practical learner selection needs regression models.

It is argued here, that it is this need to support trade-offs that defines the challenge of practical learner selection and that, as will be seen shortly, this concept is rarely, if ever, discussed in the meta-learning literature. It does not necessarily follow that other meta-learning approaches, such as ranking systems, cannot be of assistance in practical work but without providing information about what differences to expect, they must be considered of limited value.

When time may be scarce, the sequence of evaluation becomes important because evaluation times are uncertain and difficult to predict. A time constraint may not become apparent until there is insufficient time left to evaluate the best learners. But a planned sequence, an evaluation strategy, only makes sense if there is some rational basis for believing that one learner is likely to be more accurate than another. In this scenario, some of the proposed, comparative, meta-learning approaches could be of value — but only if they can achieve a required level of reliability, and there is no substantive evidence that they can.

A study of recent mainstream practitioner texts Witten et al. (2011), Tsiptsis and Chrorianopoulos (2009) and Han and Kamber (2006) reveals an absence of guidance on learner selection strategy and it may be concluded that no best-practice has emerged from the research community. The strategies used in practice are probably limited to combinations of:

- Random selection
- Prior beliefs
- Sample testing
- Exhaustive testing

Experienced-based prior beliefs about relative performance do, arguably, represent a rational basis for deciding upon an evaluation strategy, despite the No Free Lunch Theorems mentioned in the Introduction, which, it has been argued, lack practical relevance (Giraud-Carrier and Provost (2005)). Such beliefs, though, are difficult to quantify. Sample testing faces the same problem; there is no rational basis for any particular strategy. It is contended that learner selection is currently an ad-hoc process.

To progress, we need methods for numerically predicting accuracy and evaluation time; they will need explanatory variables and calculating their values must represent a good use of the available time. To provide confidence in decision making, we must be able to show that these predictions are reliable; at least approximate knowledge of their error distribution is needed.

Predictions could be used in decision-making systems to determine optimal evaluation strategies. Smith-Miles (2008) discusses how learner selection can be viewed as an instance of the more general algorithm selection problem articulated by Rice (1976) but it is debatable whether that conceptual framework is appropriate, given the need to use distributional information in the decision-making process and to select a portfolio rather than a single learner.

Determining optimal evaluation strategies may be seen as the ultimate goal in learner selection research but there is an extensive middle ground between there and the current ad-hoc approach. Reliable predictive models would allow practitioners to formulate better informed evaluation strategies. This would represent an advance towards a principled learner selection protocol. With the need to weigh the other qualitative criteria for selection, the uncertainty surrounding model predictions and often small differences between learner accuracies, it is not certain that an optimal decision-making framework would produce results that were better, in any practical sense, than a robust model-based system supporting greedy decision making.

Fig 2.1 summarises the key idea expressed in this section.

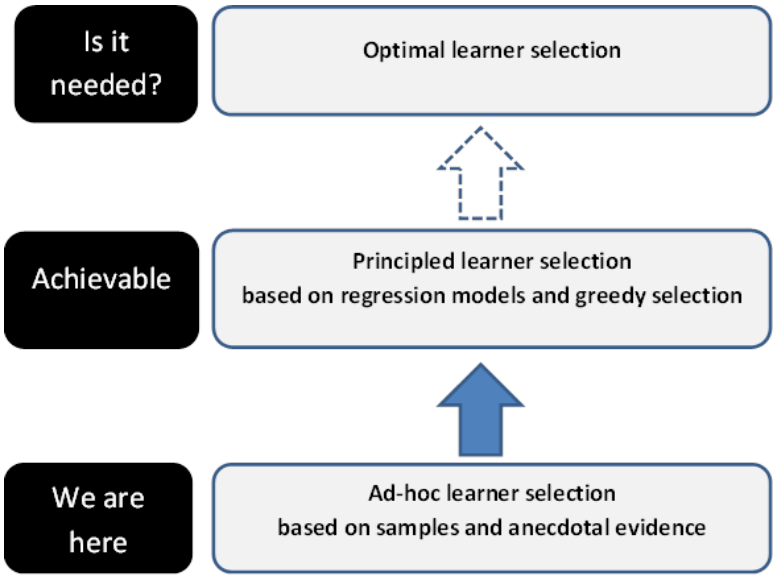

Figure 2.1: Towards principled learner selection

2.2 Predicting accuracy

In the last section a requirement was established for a process for numerically predicting the relative accuracy of a set of learners on a dataset, without the need to fully evaluate all of them. A review of the literature reveals that substantial research effort has been expended trying to achieve *similar* aims, with the work naturally dividing into two high-level categories:

1. Direct sample-based approaches
2. Model-based approaches

Direct sampling is a straightforward process:

- Evaluate a learner on a sample of the data
- Use the accuracy as an estimate of the accuracy on the full dataset
- Extrapolate the full evaluation time from the sample time

In a model-based (meta-learning) approach, a relationship is sought between the response variable, accuracy or time, and a set of explanatory variables derived from the dataset. Sample estimates may also be used as explanatory variables in a modelling process. Each of these approaches will now be considered separately.

2.2.1 Sample-based approaches

The focus in these studies has been on answering questions like:

1. Should samples be drawn statically or dynamically? (see John and Langley (1996))

2. Is the use of progressively larger sample sizes more efficient than using single fixed sizes? (see Provost et al. (1999))

These questions are not germane to this work, where a fixed sampling strategy was sufficient to achieve the experimental aims. Whilst future work involving sample estimates in models may draw upon the results of such papers, our interest here is because they provide a benchmark against which to assess meta-learning. For example, Provost finds that geometric sampling was reported as being between three and thirty times faster than learning with all the data and that the error rate on the final sample was within 0.005 of that on all the data. This is a level of accuracy an order of magnitude beyond any reported in the meta-learning literature.

Petrak (2000), perhaps uniquely, acknowledged that sampling and meta-learning research represent two different paths to predicting learner accuracy. This paper investigated how well the most accurate learner for a specific dataset could be predicted using a non-progressive (fixed) sample strategy. A simple 1000 train/1000 test protocol was found to use 54 times less evaluation time than 10-fold CV, with the most accurate of 8 learners being selected 22 out of 35 times. Similar results are presented in this work.

Smith et al. (2012) used regression models to predict error-rates from samples using the bias-variance decomposition of the sample as the explanatory variables. The evidence that this approach is any better than modelling using the sample estimate was inconclusive. As in Petrak (2000), 1,000 is seen to be a sample size that supports good predictive results.

An explicit, quantified, acknowledgement of the relationship between the trade-off between predictive accuracy and the time that could be being used to train learners is a feature of the sampling literature but is not seen at all in meta-learning papers. The simplicity of these sample approaches provides a further contrast with the often complicated, non-sample oriented approaches, that will be discussed shortly.

2.2.2 Model-based approaches

The template followed by many meta-learning studies was arguably set by the published results of the Statlog project (Michie et al. (1994)). 16 general char-

acteristics of datasets were considered. 5 were structural measures such as the number of instances; 6 were statistical measures such the mean absolute correlation of attributes. The remaining 5 were based on information theory, such as the mean entropy of the attributes. A set of classification rules was induced by another learner, C4.5.

Here is one of the rules:

$$CART\ appl\ if\ N \leq 6435, Skew > 0.57$$

This rule states that the CART learner is 'applicable' if the number of instances is below a threshold value whilst mean attribute skewness is above another empirically established threshold. By applicable, it is meant that for datasets with the qualifying number of instances and level of skewness, CART delivered a classification accuracy within 8 standard deviations of the lowest error rate achieved on that dataset by any of the 23 tested learners. This type of rule does not fit the requirements for contributing towards the concept of principled learner selection, which requires numerical prediction of accuracy. These same SSI measures were used in many subsequent studies, even though the Statlog results provided only weak evidence of their efficacy.

Sohn (1999) used the Statlog data in one of the few published studies to attempt regression modelling of learner error rates. Mean absolute prediction errors (MAE) were in the range [0.02, 0.08] (average 0.06) for a set of 11 learners.

Two other general approaches to predicting learner performance have appeared in the literature; pair-based classification and similarity-based ranking. In the former (e.g. Kalousis (2002)), classification rules are induced for predicting which of a pair of learners would perform best on a dataset. In the latter (e.g. Brazdil and Soares (2000)), a nearest-neighbour learner uses the explanatory variables to identify previously examined datasets that have similar characteristics to the one under test. The relative learner performances on these similar datasets are then used to produce a set of predicted rankings for the test set.

Pfahringer et al. (2000) crossed the boundary between sample-based and non-sample approaches by introducing the concept of a landmark score. This is the error rate of a learner that has a relatively fast training time, the idea being that it will be a useful predictor of how slower but more sophisticated and (generally) accurate learners may perform on the same data. It is suggested in the paper that explanatory variables should not exceed O(n log n) complexity in order not to waste time calculating metrics that could be used to train learners. A linear discriminant (LDA) classifier, Naïve Bayes and a C4.5 learner were used as landmarks with 10-fold CV employed as the evaluation protocol on the full

datasets. The results were inconclusive but suggested that the classification rules were more useful than guessing.

Fürnkranz and Petrak (2001) extended the use of sample-based explanatory variables. They used the ratios of landmark scores and their relative rankings in addition to their absolute error rates as predictors, recognising the possibility of there being important performance interactions between different landmarks. They also evaluated more sophisticated learners on sub-samples of fixed size 100 and 200 (minimum dataset size 1000). The re-substitution error was used rather than 10-fold CV as an evaluation protocol. No non-sample explanatory variables were used. All results were interpreted negatively.

Peng et al. (2002) proposed a new series of 15 explanatory variables based on measures derived from decision trees induced on each dataset. No significant difference in performance between models using these variables or ones using the SSI variables was detected.

Ho and Basu (2002) introduced a set of measures aimed at defining the geometrical complexity of a dataset. Whilst ostensibly developed to assist in the study of learner performance (e.g.Mansilla and Ho (2005) / Luengo and Herrera (2010) / Trujillo et al. (2011)), they nonetheless stand-out as potentially strong explanatory variables for meta-learning work. The authors remark that whilst earlier work had been based on statistical and information theoretic measures, 'in classification, it is the geometry that counts most'. In fact, a number of the measures could be categorised as landmarkers or grouped alongside the existing bank of statistical or structural measures. The measures designated with the codes N1 and T1, though, undoubtedly measure aspects of classification boundary complexity that previously used statistical measures do not. They have yet to feature in a predictive study.

Two, more recent, regression studies (Abdelmessih (2010), Reif et al. (2012)) modelled error rates using SSI and landmark variables, with Reif also employing the tree-based variables proposed by Peng. Neither employed complexity measures. An important difference between these studies was that only Reif employed a feature selection process. The best results for Abdelmessih were obtained using only the landmark variables and were in the range [0.05, 0.08] (average 0.07) for a set of 7 learners. Reif returned similar results for landmark only models and slightly better results with models with all variables presented to the feature selection process. The feature selected models performed much better than the full models without selection in Abdelmessih. For both studies, the SSI sets performed much worse than landmark only sets, with MAEs almost double those for the landmark set. It is likely, in view of the poor performance of landmarks observed in this work, that the apparent relative success of landmark variables in

these studies may be attributable to the fact that the set is much smaller than the SSI set, reducing the tendency to overfit, rather than to superior predictive power.

2.3 Predicting evaluation time

There are few published papers that consider the prediction of learner evaluation time. Whilst sample-based studies generally mention sample and total evaluation times, their interest does not extend to modelling the relationship between the two. Lindner and Studer (1999) proposed a case-based reasoning system that allowed the user to express a broad evaluation time requirement, for example that a 'very fast' learner was required. Brazdil and Soares (2000) combined accuracy and prediction time into a single ratio with a weighting to reflect the trade-off that a practitioner would accept. Ali and Smith (2006) follow a similar approach. None of these methods are of interest here.

The only study reviewed that attempted regression of evaluation time was Reif et al. (2011). 34 simple, statistical, information theoretic and tree-model measures were used as explanatory variables along with the computation times for each of these broad categories (i.e. the time taken for computing all the statistical attributes). The Naïve Bayes, one-nearest neighbour and decision stumps learners were used as landmarks; their training times used in the modelling process. It was found that feature sets with landmark times performed better than those without although the difference in performance was not significant.

2.4 Conclusions

We have seen two contrasting approaches to predicting learner accuracy. Direct sample approaches arguably offer the following advantages over model-based approaches:

- Sample protocols are easier to implement than modelling processes
- The interpretation of how results are derived is obvious, engendering confidence in them
- There is a clear and controllable relationship between the processing time required to get predictions and their resulting accuracy (via control of the sample size)
- Predictions are numerical values, as required for principled learner selection

Few meta-learning studies attempted regression of learner performance but the available results, when viewed in the light of the analysis of mean absolute errors to be presented here, is poor. It is also noteworthy that the use of landmarks (samples) was felt to improve upon the performance of SSI variables, although it was remarked here that this effect may be an artefact of feature selection.

The case for pursuing research into model-based approaches, using non-sample variables, appears weak but there is an important gap in our knowledge — no study has evaluated SSI and landmark variables alongside the Ho and Basu complexity measures, nor has one used sample estimates from sophisticated learners as explanatory variables.

Table 2.1 lists 25 meta-learning studies showing the range of explanatory variables used by each, highlighting this absence of a comprehensive regression study.

study	datasets	synthetic	learners	statinfo	complexity	landmark	sample	tree	time	regression
Aha (1992)	1	✓	4	✓						
Michie et al. (1994)	22		23	✓						
Sohn (1999)	19		11	✓						✓
Lindner and Studer (1999)	80		21	✓						
Pfahringer et al. (2000)	×	✓	3	✓		✓				
Bensusan and Giraud-Carrier (2000)	17	✓	10	✓		✓				
Köpf et al. (2000)	×	✓	3	✓						
Brazdil and Soares (2000)	16		6	✓					✓	
Petrak (2000)	35		8				✓			
Fürnkranz and Petrak (2001)	48		5			✓	✓			
Kalousis (2002)	65	✓	8	✓		✓				
Peng et al. (2002)	47		10	✓		✓		✓		
Ho and Basu (2002)	14	✓	×		✓					
Singh (2003)	10		4		✓					
Kalousis et al. (2004)	80		10	✓						
Leite and Brazdil (2005)	30		2	✓			✓			
Mansilla and Ho (2005)	14		5		✓					
Ali and Smith (2006)	100		8	✓					✓	
Lee and Giraud-Carrier (2008)	135	✓	7	✓						
Cacoveanu et al. (2009)	13		9	✓			✓			
Abdelmessih (2010)	90		7	✓		✓				✓
Brazdil et al. (2010)	80		10	✓			✓			✓
Macia (2011)	70		3		✓					
Reif et al. (2011)	34		5	✓		✓		✓	✓	✓
Reif et al. (2012)	54		9	✓		✓		✓		✓
This study (2013)	**50**	✓	**10**	✓	✓	✓	✓	✓	✓	✓

Table 2.1: Twenty-five meta-learning studies 1992-2013

It is perhaps surprising that so few papers have been published on predicting evaluation time, with only one regression study reviewed. Either it has not been recognised as being, arguably, as important as accuracy in learner selection, or the results were too disappointing to be published (publication bias).

The meta-learning literature presents no analysis of the time efficiency of the proposed approaches. For example, is it worth calculating SSI variables for a dataset, or should that time be spent evaluating a learner? In a similar vein, there is a notable absence of any analysis on the practical significance of the reported results. For example, how does a mean absolute accuracy of 7% for a regression model translate into the distribution of errors a practitioner might expect?

2.5 Summary

A review of related work in the area of learner performance prediction has revealed the following areas in which contributions could be made:

1. A model-based study using all categories of explanatory variable
2. An analysis of the trade-off between evaluation time and predictive accuracy for model-based approaches
3. An analysis of the practical significance of the mean error rates associated with model-based approaches

Chapter 3

Experimental design

In the Introduction, the viability of meta-learning was questioned. Chapter 2 expressed the ambition of advancing to principled learner selection from the current ad-hoc approach and it was contended that regression models would be needed to support the informed trade-offs that are central to the concept. In this chapter, starting from the knowledge gaps in the literature identified in Chapter 2, we progress to an experimental design. In Section 1 we set out the rationale behind the experimental work, leading to the formulation of a set of experimental aims. Section 2 explains the experimental strategy, before some of the more important choices are discussed in greater detail in Section 3. Section 4 provides a detailed account of each stage in the procedure.

3.1 Experimental aims

The regression studies reviewed in the previous chapter announced mean absolute predictive errors of between 6% and 12%, depending on the set of explanatory variables used. Preliminary analysis for this work revealed that the average difference between the 1st and 8th ranked learner by accuracy on a set of 50 diverse datasets was around 5%, suggesting that the attainable level of predictive accuracy with SSI and landmark variables is too low to differentiate between learners and hence could not support a move towards principled learner selection.

The literature review also revealed that no meta-learning experiment has used the Ho and Basu complexity measures alongside the frequently tested SSI and landmark measures or used them in a regression model of learner accuracy. If complexity measures cannot increase the accuracy of predictive models, then we might conclude that non-sample explanatory variables should be ignored in further research towards principled learner selection. To draw such a conclusion, we would need a more precise analysis of model accuracy, with distributional intuition into what MAE means to a practitioner.

Furthermore, no experiment has used sample estimates as explanatory variables alongside all of these other potential predictors. This is pertinent because it is possible that sample estimates could be enhanced by model interactions with SSI, landmark and complexity measures. Non-sample explanatory variables cannot be dismissed before this synergy has been discounted.

Alternatively, it may be the case that sample-estimates interact with each other in modelling processes, producing a higher level of accuracy than they would be capable of achieving singularly. For example, if a combination of samples in a model could predict the accuracy of another learner, not sampled, then there is the potential to use evaluation time more efficiently.

Acting as a backdrop to this desire to determine exactly how much predictive power non-sample variables have, is the question mark over their time efficiency. Michie et al. (1994), Pfahringer et al. (2000) and Fürnkranz and Petrak (2001) all questioned whether time spent calculating these variables would be better spent on evaluating learners but no-one has attempted to provide an evidenced answer.

The various questions posed in this section can be condensed into the following two experimental aims:

1. To determine if accuracy models need non-sample explanatory variables
2. To determine if there is a synergy between sample-estimates that modelling can exploit

Finally, it was decided to use the same experimental set-up to explore the modelling of evaluation time. As the literature provides few details of what may be expected, no firm experimental aims were set for this phase.

3.2 Experimental strategy

We want to develop regression models to reliably predict the accuracy of a learner on a previously unseen dataset. To apply the regression model, the explanatory variables used in the model must be calculated for the dataset; this takes time. Thus each model has two quantifiable characteristics; an accuracy level and a processing time, both of which are determined by the combination of explanatory variables used. If two models offer the same level of accuracy then the one with the lower application time will be preferred.

In theory, one model could be built per learner for every possible subset of the set of explanatory variables and the accuracy and application time of each recorded. With 57 non-sample variables plus dozens of potential sample-based

variables, the power-set of variables is astronomically large and so exhaustive testing is not viable. The compromise used here is to use the groupings by which the variables were introduced in the literature as de facto variables, with each group included or excluded in its entirety in the model building process. These groupings, with membership counts, are as follows:

1. Structural (8)
2. Statinfo (15)
3. Tree-based (4)
4. Landmarks (5)
5. Complexity (14)
6. Proposed (9)
7. Complexity lite (5)

Table 3.1 lists each variable individually, with brief details and references to further information.

To determine whether non-sample variables are needed in accuracy modelling we add a group to a sample-based model and note the change in model accuracy. This change can then be assessed in light of the additional model application time required for the extra variables to be calculated. Again, there is a need to restrict the experiment to only a few sample strategies and these are listed in Table 3.2.

To establish some benchmarks, each non-sample group and sample strategy will first be tested in isolation. Even using this block approach there are hundreds of potential combinations of groups and sampling strategies, each of which can be considered as a distinct modelling strategy. Only 20 will be tested but the results suggest that this guided selection is adequate for providing the answers sought.

Each learner will have a regression model built for each modelling strategy. Different modelling strategies cannot be expected to retain their relative accuracy across all learners but rather than try to analyse the performance of a strategy at learner level, which would increase the effective number of modelling strategies by a factor equal to the number of learners, it seems sensible to consider them as a 'system'. So, the interest is in how a modelling strategy impacts on the overall prediction error of the group of learners.

The measure of accuracy of a modelling strategy will be the System Mean Absolute Error (sys mae), which is the sum of the MAEs for each learner when trained and tested, using a 10 x 10 CV process, with the variables associated with that strategy, over the pool of datasets used in the experiment. The MAE

identifier	group	detail
inst	struct	number of instances
attr	struct	number of attributes
maj	struct	% of the majority class
nom	struct	number of nominal attributes
num	struct	number of numeric attributes
missVal	struct	% of missing values
missAttr	struct	% of attributes with missing values
missInst	struct	% of instances with missing values
sdr	statinfo	standard deviation ratio (Michie et al. (1994) pg.115)
boxM	statinfo	Box's M statistic (discretized)
cancor	statinfo	canonical correlation
hotel	statinfo	Hotelling's T-statistic (discretized)(Köpf et al. (2000) pg.5)
lda	statinfo	resubstitution error of linear discriminant learner
intercor	statinfo	average correlation between attributes
maxcor	statinfo	maximum attribute-class correlation
avecor	statinfo	average correlation between attributes & class
skew 12	statinfo	univariate ave. attribute skewness by class
kurt 12	statinfo	univariate ave. attribute kurtosis by class
related	statinfo	proportion of attrs with chi-sq association to class
entropy	statinfo	average entropy per attribute (Michie et al. (1994) pg.116)
mutual	statinfo	average mutual information between attrs &class
nsratio	statinfo	noise to signal ratio
enatrr	statinfo	effective number of attributes
f1v	complex	directional-vector max. Fishers discriminant ratio (Macia (2011) pg.28)
f1	complex	Fisher's discriminant ratio (Ho and Basu (2002))
f2	complex	volume of overlap region
f3	complex	feature efficiency
f4	complex	collective feature efficiency (Macia (2011) pg.30)
l1	complex	minimised error by linear programming
l2	complex	error rate of linear classifier by linear programming
l3	complex	nonlinearity of linear classifier by linear programming
n1	complex	proportion of insts on class boundary (MST method)
n2	complex	ratio of ave. inter/intra class nearest neighbour dist.
n3	complex	error rate of 1NN classifier
n4	complex	nonlinearity of 1NN classifier
t1	complex	proportion of insts with adherence subsets retained
t2	complex	average insts per attribute
treeHW	tree	decision tree height-to-width ratio (Peng et al. (2002))
treeNH	tree	decision tree nodes-to-height ratio
treeLW	tree	decision tree length-to-width ratio
treeHP	tree	decision tree height-to-minimum path length ratio
l-lda	landmark	66% hold-out linear discriminant error-rate (Pfahringer et al. (2000))
l-knn	landmark	66% hold-out 1-nearest neighbour error-rate
l-oner	landmark	66% hold-out Holte one-rule error-rate (Holte (1993))
l-nbay	landmark	66% hold-out naive bayes error-rate
l-stump	landmark	66% hold-out decision stump error-rate
noise	proposed	prop. of unique attribute tuples with different class labels (clashes)
overlap	proposed	prop. of insts closer to the other class than their own (Mahalanobis dist.)
outliers	proposed	prop. of insts with Mahalanobis dist from class mean beyond 5% critical value
clusters	proposed	number of clusters identified by EM algorithm (200 sample)
clusprop	proposed	prop. of insts in smaller of EM fixed 2-clusters (200 sample)
bayratio	proposed	ratio of naive bayes accuracies continuous/discretised
lkratio	proposed	ratio of landmark accuracies l-lda / l-knn
mnorm	proposed	Mardia test of multivariate normality on 200 sample (Mardia (1970))
minval	proposed	minimum Chi-squared attribute-to-class p-value

Table 3.1: The 57 tested explanatory variables

strategy	detail
smp_a	10 x 10 CV 25% sample per dataset - all learners
smp_b	10 x 10 CV 50% sample jrip/j48/forest/knn/logistic
smp_c	10 x 10 CV 100% oner/nbayes/bayesnet + 50% jrip/j48/forest/knn/logistic
smp_d	1 x 10 CV per dataset - all learners
smp_e	10 x 66% holdout per dataset - all learners
smp_f	1 x 66% holdout per dataset - all learners

Table 3.2: Sampling strategies

for each learner is the average magnitude of the differences between its predicted error-rate for a test dataset and the learners's actual error-rate on that dataset as ascertained by 10 x 10 CV.

The processing time for each modelling strategy is the total time taken to calculate the values for all of the variables, sample and non-sample, required for the strategy, across the pool of datasets used in the experiment. To help assess the relationship between accuracy and time, a measure called efficiency will be calculated for each modelling strategy:

$$efficiency = \frac{improvement\ in\ system\ mae}{proportion\ of\ evaluation\ time\ used}$$

The improvement is that between the achieved MAE and the default MAE value. The default value is obtained by using a learner's average error-rate as the model prediction for each test dataset. The evaluation time is the total time required to obtain 10 x 10 CV estimates for each learner on all datasets in the pool.

3.2.1 Strategy summary

Here is a summary of the new terminology introduced in this section:

- Variable group – a set of non-sample variables related by first literature appearance e.g. SSI variables or complexity measures
- Sampling strategy – a combination of learners evaluated on samples of different sizes from the dataset(s) e.g. all learners tested on a 10% sample or learner 1 tested on 50% and the rest on 25%
- Modelling strategy – a combination of a sampling strategy and variable group e.g. smp_a + SSI
- Learner system – the set of available learners
- System MAE – the sum of the MAEs of each learner model when tested (10 x 10 CV) on the meta-set

Fig 3.2 (at the end of the chapter) attempts to assist with visualising the proposed strategy.

Each learner is evaluated in turn. The 'modelling strategy' in the diagram may be considered as the training set (meta-set) presented to the meta-learner, consisting of one instance per dataset, with attributes as per the strategy (e.g. SSI variables) and the error-rate for that learner on those datasets as the concept to be learnt.

Each variable group and sample strategy are timed separately to the accuracy evaluation, over all the evaluation datasets as a block. The overall time for a modelling strategy is then paired with the overall system MAE. The experimental procedure, detailed in section 1.4 below, should help clarify matters further.

3.3 Experimental choices

This section discusses some of the key parameter decisions taken for this experiment.

3.3.1 The measure of accuracy chosen

The limitations of accuracy (or its compliment, error-rate) as a comparator are well known; alternatives exist. Bradley (1997) and Provost et al. (1997) argue for the use of AUC as a single number metric, although Hanczar et al. (2010) cautions against using AUC unless the sample sizes are very large. Ben-David (2008) makes a case for the Kappa score, demonstrating that it offers many of the benefits of AUC with the advantage of being easier to calculate and extend to m-class problems.

The interest here is with evaluating modelling approaches rather than deciding whether one learner is more useful than another, so it is questionable whether the concerns about error rate directly apply. However, as the aim is to develop practical tools for domains in which error rate may not be appropriate, we should, ideally, ensure that experimental findings are consistent across measures. This would increase the workload significantly. As the various measures are highly correlated with each other it seems reasonable to believe that factors that explain the variation of one will also explain the variation of the others to a similar degree and so the extra workload is difficult to justify. This work uses error rate as the target measure for two reasons:

1. It enables comparison of results with similar studies

2. There is no unarguable case for not doing so

3.3.2 The use of 10 x 10 cross-validation

There are a number of re-sampling methods for estimating off-training set (OTS) error but stratified 10-fold cross-validation is considered to represent the best balance between computational overhead and estimate reliability (Kohavi (1995)) when data is limited. Repeating the process 10 times and using the average of the results as the final estimate is a recommended best practice for achieving a reliable result (Witten et al., 2011, p. 154).

3.3.3 Deciding how many datasets to use

Crawley (2005)[p. 9] advises that, generally 'a sample of 30 or more is a big sample, but a sample of less than 30 is a small one'. For regression modelling, he states as a rule of thumb (p. 204) that a maximum of N/3 explanatory variables (EVs) should be fitted during a multiple regression, where N is the number of observations. Green (1991) provides a survey of opinions on the matter from which it was concluded that a minimum of 5 observations per EV are required but that 10 or more would be preferable. As, after feature selection, most models are presented with 5-10 EVs, fifty seemed a reasonable number of datasets to use. Macia (2011)[p. 46] presents an empirical justification for a test-bed size for machine learning comparisons of between 20 and 150 datasets.

3.3.4 Selection of datasets

As some explanatory variables can only be calculated for 2-class scenarios, it was decided to only use 2-class datasets. In order to enable consistent evaluation of statistical variables, it was decided that each dataset should have at least 3 continuous attributes.

Approximately, half of the datasets were sourced through the Dcol library. These datasets were originally from the UCI repository, with m-class problems having been converted to 2-class problems by discriminating one class against another. Table 3.3 lists the repositories used and Table 3.4 provides structural details of the 50 datasets selected, an asterisk indicating that a regression problem was converted to a classification. Further details of the domains are provided in Appendix A.1.

Referring to the UCI repository, Holte (1993) concluded that 'most of these datasets are typical of the data available in a commonly occurring class of 'real' classification problems', although Salzberg (1997) cautions that 'the UCI repository is a very limited sample of problems, many of which are quite easy for a classifier'. The repository has grown very significantly since those statements,

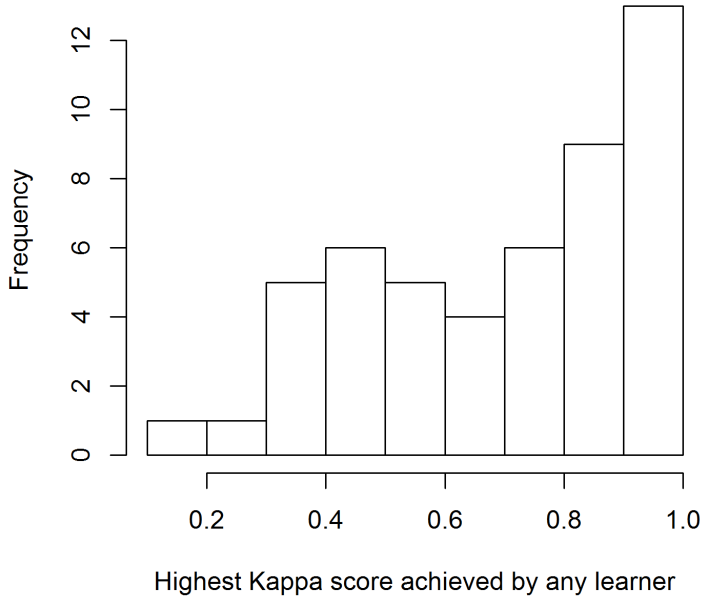

Figure 3.1: Highest Kappa score distribution of the datasets

both in respect of the size of datasets available and the range of domains from which they are drawn. Still, Macia et al. (2012) identified gaps in the UCI coverage when dataset complexity is considered.

Until a comprehensive collection of test sets is agreed by the research community, the only sensible strategy is to select from as wide a range of domains as possible, to ensure a reasonable range of problem difficulties and to be cautious in generalising the results beyond the collection. In assessing the range of problem difficulties, looking only at the range of average or maximum classification accuracies can be miss-leading as class balance can be a highly influential factor. A suggestion made here is to instead use the highest Kappa score attained by any of the learners on each dataset as a measure of classification difficulty. Fig 3.1 shows the distribution of these scores for the 50 datasets used here, illustrating a reasonable spread across the potential range.

Another problem faced by machine learning researchers is that many of the repository datasets are tiny in comparison to some of the datasets faced by practitioners. With hindsight, more time should have been invested in finding larger datasets. The size issue is much more important when modelling evaluation time, which is very noisy, a fact not fully appreciated when the collection was compiled.

short	title	reference
dcol	Data Complexity Library in C++	Orriols-Puig et al. (2010)
UCI	UCI machine learning repository	Frank and Asuncion (2010)
statlib	Statistics, Carnegie Mellon University	
Keel	KEEL-dataset repository	Alcal-Fdez et al. (2011)

Table 3.3: Details of the public repositories used for dataset provision

3.3.5 Selection of learners and meta-learners

The WEKA toolkit (Witten et al. (2011)) was used for the experiment. 10 learners (Table 3.5) were selected from across the available algorithm categories (rules, functions, lazy etc.) in order to assemble a learner system with a range of different biases. No pre-processing techniques were used for the base-level classifications.

For each modelling strategy the WEKA linear regression and SMO regression learners were both used to build models. Because of the large sets of explanatory variables being presented to the meta-learner, it was decided to employ a feature selection process. (Witten et al., 2011, pg.308) discusses the negative impact of irrelevant attributes on learner performance. The results of Reif et al. (2012) show that feature selection can improve model accuracy significantly. Correlation feature selection (cfs) was employed and the filtered attribute set was presented to both meta-learners. The linear regression learner also used the M5 attribute selection process. The aim for the meta-learning stage is to deliver the best model possible given the presented variables, so it is sensible to use this facility where available (linear regression) even if an equivalent procedure is not available for SMO regression. It is not an experimental aim to compare the two meta-learners.

Use of the multi-layer perceptron was considered but ad-hoc testing suggested it was consistently slower and less accurate than the other two regression learners and so its use was discounted.

3.3.6 Sampling protocols

Instances were drawn at random but the number for each class was determined in advance so that the sample was coerced into having the same class distribution as the full dataset. This approach could perhaps be described as simple stratification. It was felt that this would produce a more representative sample than a pure random selection, especially when employing percentage-based sample sizes with smaller datasets.

In order to use sample estimates in models, consistent sample strategies must be employed, regardless of differing dataset sizes. In practice, a range of models would be required to accommodate different sample sizes, so that the available

id	source	inst	attr	nom	num	maj (%)	missval (%)	accuracy (%)
100	dcol - adl	48,842	14	8	6	76.1	0.95	83.9
101	dcol - authors	841	70	0	70	62.3	0.00	97.2
102	dcol - bal	625	4	0	4	53.9	0.00	87.8
103	dcol - bpa	345	6	0	6	58.0	0.00	62.5
105	dcol - cmc	1,473	9	7	2	57.3	0.00	66.8
106	dcol - col	368	22	15	7	63.0	23.80	82.2
107	dcol - crx	690	15	9	6	55.5	0.65	84.6
108	dcol - drm	366	34	1	33	69.4	0.06	99.4
109	dcol - ecu	736	19	5	14	75.5	3.20	91.6
110	dcol - h-s	270	13	0	13	55.6	0.00	80.7
111	dcol - ion	351	33	0	33	64.1	0.00	87.7
114	dcol - mag	19,020	10	0	10	64.8	0.00	80.5
116	dcol - opt	5,620	62	0	62	90.1	0.00	99.1
117	dcol - pbc	5,473	10	0	10	89.8	0.00	94.9
118	dcol - pen	10,992	16	0	16	89.6	0.00	98.3
119	dcol - pim	768	8	0	8	65.1	0.00	74.9
120	dcol - seg	2,310	18	0	18	85.7	0.00	97.7
121	dcol - spa	4,601	57	0	57	60.6	0.00	89.1
123	dcol - veh	846	18	0	18	74.9	0.00	76.0
125	dcol - wav21	5,000	21	0	21	66.9	0.00	84.9
126	dcol - wbcd	569	30	0	30	62.7	0.00	94.5
127	dcol - yea	1,484	8	0	8	68.8	0.00	70.3
128	uci - vertebral	310	6	0	6	67.7	0.00	80.6
129	uci - ilpd	583	10	1	9	71.4	0.07	67.8
130	uci - blood	748	4	0	4	76.2	0.00	76.4
131	dcol - ann	898	30	24	6	54.1	0.00	97.0
132	uci - mammographic	961	5	0	5	53.7	3.37	80.9
133	uci - steel	1,941	33	6	27	65.3	0.00	91.2
134	uci - cardiotocography	2,126	22	0	22	77.8	0.00	91.3
135	uci - insurance	5,822	85	0	85	94.0	0.00	91.3
136	uci - bank	4,521	16	9	7	88.5	0.00	88.8
137	uci - statlog	43,500	9	0	9	78.4	0.00	97.7
139	author - carpet sales*	1,242	9	6	3	75.0	0.00	79.8
140	uci - housing*	506	13	0	13	75.5	0.00	88.2
141	uci - mpg*	398	7	0	7	50.3	0.22	86.1
142	uci - auto*	205	24	10	14	50.2	1.12	89.2
143	uci - computer*	209	8	1	7	50.2	0.00	90.5
144	uci - solar*	1,066	11	8	3	59.6	0.00	73.2
145	uci - concrete*	1,030	8	0	8	75.0	0.00	85.9
146	uci - parkinsons*	5,875	21	0	21	75.0	0.00	96.1
147	liaad - ailerons*	7,128	5	0	5	84.6	0.00	88.3
148	statlib - colleges*	1,302	31	0	31	75.0	18.35	89.4
149	statlib - houses*	20,640	8	0	8	75.0	0.00	87.1
150	statlib - irish*	500	5	3	2	69.8	0.00	86.7
151	statlib - NO2*	500	7	0	7	75.0	0.00	77.4
152	uofn - protein	23,464	7	0	7	69.8	0.00	76.5
153	keel - phoneme	5,404	5	0	5	70.7	0.00	81.0
154	keel - sa heart	462	9	1	8	65.4	0.00	69.7
155	keel - cylinder	539	19	0	19	57.9	5.38	64.9
156	keel - marketing*	8,993	13	0	13	80.6	2.30	86.8
		5,049	19	2	16	69.4	1.18	84.9

Table 3.4: Basic details of the 50 diverse datasets with mean acheived accuracy

identifier	type	notes
oner	rules	Holte 1-rule simple classifier
nbayes	prob.	Naive Bayes classifier using estimator classes
j48	tree	generates a pruned C4.5 decision tree
jrip	rules	RIPPER implementation using pruning
logistic	function	generates a logistic regression model with a ridge estimator
mlp100	function	multi-layer perceptron - uses backpropagation - 100 training epochs
svmpoly	function	support vector machine with polynomial kernel
bayesnet	prob.	Bayes Network learner using simple estimation
knn	lazy	10-nearest neighbours classifier with no distance weighting
forest	tree	constructs a forest of 10 random trees

Table 3.5: Details of the 10 WEKA implemented learners used

evaluation time could be fully utilised. Here, a decision needed to be made about the size of the entry-point sample, which would be used as a base-line against which other sample and non-sample strategies could be compared, and whether it should be of fixed size or percentage-based. The decision was influenced by the fact that half of the datasets had less than 1,000 instances and the largest only 48,842. The literature suggests that a sample of 1,000 could have good predictive power and if all the datasets were above 4,000 instances then a fixed 1,000 sample would have been chosen as the base-line sample strategy. Such a fixed size would then be viable for even the largest of datasets, whereas percentage-based samples would be liable to grow to a size that defeats their purpose. With so many smaller datasets, any fixed size would have been too small for the largest datasets in the collection. As a compromise, the entry-point sample was set at 25% for this experiment.

3.3.7 Discretization of continuous variables

The statistical variables are only calculable on continuous attributes and the information theoretic variables on discrete attributes. As the majority of the datasets have no nominal attributes, it was decided to discretize all continuous attributes in order to allow information variables to be calculated for all datasets.

There are a number of supervised and unsupervised discretization schemes from which to choose, each offering a different balance between execution time and impact on learner accuracy. Liu et al. (2002) provide a comprehensive, comparative study.

During preliminary work, a Holte discretization (Holte (1993))was trialled but was judged to be time inefficient, taking up to a minute to process each of the larger datasets. As recommended by Witten et al. (2011), it was decided to use the fast proportional k-interval method of Yang and Webb (2001), where k equals the square root of the number of instances. This unsupervised method processed

all 50 datasets in under a minute. Ad-hoc testing suggested that the entropy measures had greater predictive power using Holte but the orders of magnitude increase in processing speed makes Yang more appropriate for this scenario.

3.4 The experimental procedure

1. The classification error-rate and total evaluation time of each learner for 10 x 10 CV on each dataset was recorded

2. The classification error-rate and total evaluation time of each learner for 1 x 10 CV on each dataset was recorded

3. The classification error-rate and total evaluation time of each learner for a 10 x 66% holdout protocol on each dataset was recorded

4. The classification error-rate and total evaluation time of each learner for a 1 x 66% holdout on each dataset was recorded

5. 25% samples of each dataset were generated and the classification error-rate and total evaluation time of each learner for 10 x 10 CV on each sample was recorded

6. 50% samples of each dataset were generated and the classification error-rate and total evaluation time of 8 learners (excl. mlp100 and svmpoly) for 10 x 10 CV on each sample was recorded

7. The 14 complexity metrics for all 50 datasets were generated using the Dcol application and the total time for the process recorded

8. The 5 complexity lite metrics were calculated for each dataset using the 25% sample

9. The structural characteristics of each dataset were recorded

10. 12 statistical measures were generated for each dataset using R and the generation time recorded

11. Discretized versions of each dataset were produced using a proportional k-Interval discretization

12. 5 information theoretic measures were recorded along with the calculation time

13. 9 explanatory variables proposed in this work were calculated for each dataset using R and WEKA and the time required to calculate them recorded

14. 5 landmark estimates for each dataset were obtained using a 66% holdout and the time required to do so recorded

15. 5 tree-based measures for each dataset were obtained and the time required to do so recorded

16. Datasets (meta-sets) were prepared for regression modelling of the 10 x 10 CV error-rate obtained on each full dataset using 20 different modelling strategies. For each strategy, one dataset per learner was produced (200 datasets in total). Each of these meta-sets had 50 instances, one for each of the base-level datasets

17. For each modelling strategy, regression models were built using the WEKA linear regression and SMO regression learners using a 10 x 10 CV process. The average MAEs over the 10 runs were recorded for each learner, for each strategy

18. For each modelling strategy the 10 x learner average MAE values were combined to give an overall total MAE for the 'system' of these learners operating on these datasets

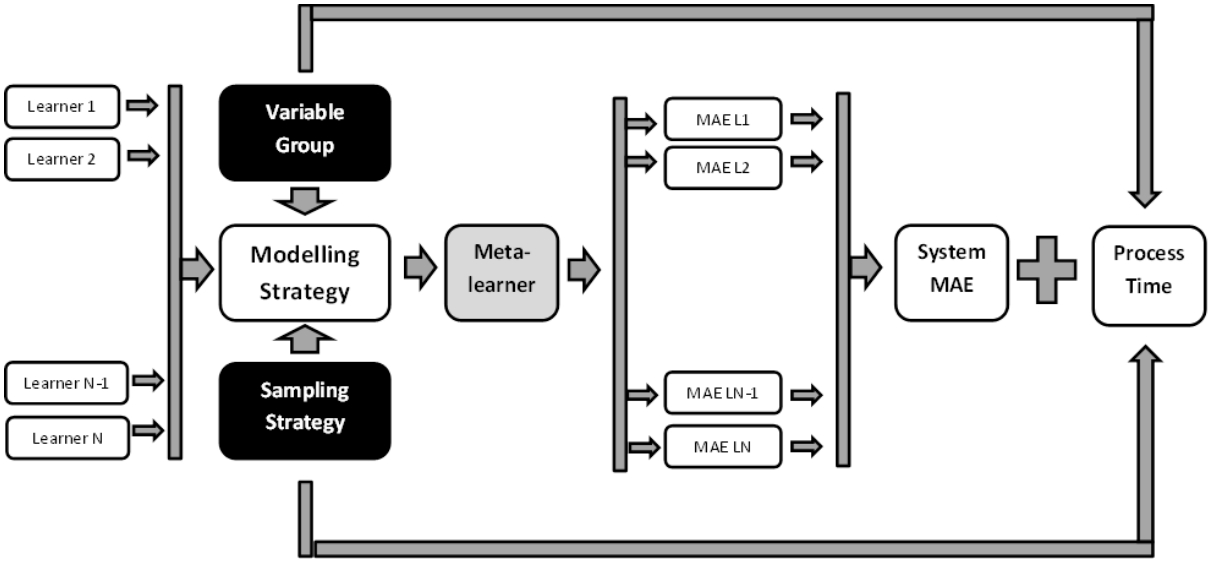

Figure 3.2: Experimental strategy

Chapter 4

Results

This chapter presents and then discusses the experimental results. After a re-statement of the aims in Section 1, Section 2 presents the results of modelling learner accuracy. The results of modelling evaluation time are presented in Section 3. Section 4 discusses, separately, the results for accuracy and time. Section 5 summarises the chapter.

4.1 Aims

In Chapter 3, two aims for the experimental work were formulated. They are re-stated here:

1. To determine if accuracy models need non-sample explanatory variables
2. To determine if there is a synergy between sample-estimates that modelling can exploit

4.2 Results accuracy

4.2.1 Base-level data

Appendix A (tables A3 – A9) lists the classification error-rates and explanatory variable values for the 50 datasets.

4.2.2 Sample-estimate times

Table 4.1 presents the total evaluation time required to produce each of the sample estimates that will be used in the five sampling strategies presented in Chapter 3. It shows, for example, that creating a sample estimate on a 50% sample of each of the 50 datasets took 1,767 seconds for the *jrip* learner but only

learner	samp 25%	samp 50%	1 x 66% hold	10 x 66% hold	1 x 10 CV	10 x 10 CV
oner	172	254	4	30	38	403
nbayes	102	186	5	10	21	296
j48	231	422	30	257	144	1,377
jrip	1,267	1,767	77	968	613	5,970
logistic	771	1,106	46	469	226	2,216
mlp100	35,280	×	1,573	15,571	6,241	66,001
svmpoly	6,731	×	1,676	8,307	11,900	140,362
bayesnet	232	376	5	52	48	502
knn	803	1,784	576	5,928	835	8,279
forest	786	799	54	501	226	2,418
total	46,375	6,695	4,046	32,092	20,292	227,824

Table 4.1: Evaluation times (secs) for the fifty datasets by learner

186 seconds for *nbayes*. These times are obtained from the raw user evaluation training and testing time (per fold) data recorded by WEKA. Experiments were run on a personal computer with 2.3 Ghz processor, 8 Gb RAM, using a 64-bit operating system (Windows).

4.2.3 Sample-strategy times

Table 4.2 presents the processing times for each sample strategy, calculated using the data in Table 4.1.

If we were to be presented with these 50 datasets and we wanted to predict the 10 x 10 CV error-rate for each of them for each of the 10 learners, electing to use 10 regression models (one per learner) developed around sampling strategy B (smp_b), we would be required to perform 10 x 10 CV using *jrip*, *j48*, *forest*, *knn* and *logistic* on 50% samples of all 50 datasets, in order to produce the required explanatory variables. Summing the appropriate values from Table 4.1, we see that smp_b has an evaluation time of 5,879 seconds, compared to a full evaluation time of 227,824 seconds, a time saving of 97.4%.

strategy	detail	seconds
smp_a	10x10 CV 25% sample per dataset - all learners	46,375
smp_b	10x10 CV 50% sample jrip/j48/forest/knn/logistic	5,879
smp_c	10x10 CV 100% oner/nbayes/bayesnet + 50% jrip/j48/forest/knn/logistic	7,080
smp_d	1x10 CV per dataset - all learners	20,292
smp_e	10x66% holdout per dataset - all learners	32,092
smp_f	1x66% holdout per dataset - all learners	4,046

Table 4.2: Sampling strategies with evaluation times

4.2.4 Non-sample variable group times

Table 4.3 presents the processing times for each group of non-sample explanatory variables, across all 50 datasets.

group	seconds
structural	5
statinfo	40
tree–based	25
landmarks	180
complexity	23,580
proposed	250
complexity lite	1,080
total	25,160

Table 4.3: Calculation times for different groups of non-sample explanatory variables on the 50 datasets

4.2.5 Main result - modelling strategy performance

Table 4.4 presents the accuracy versus time analysis for 20 selected modelling strategies. Column 1 states the total MAE for the system of 10 learners on these datasets as determined by 10 x 10 CV processes. Column 2 states the MAE improvement as a percentage of the default MAE (0.8384). Columns 3 and 4 give the processing time for the strategy and the percentage of the full evaluation time that it represents. Column 5 gives the calculated efficiency measure defined earlier. The table has been ordered by system mae, column 1.

If we wanted to use regression models based on the estimated error-rates of each learner on a 25% sample, plus the complexity measures, our modelling strategy would be smp_a/complexity and using tables 4.1 and 4.3 we see that a processing time of 69,549 seconds or 30.5% of the full evaluation time would be required. Note that this particular modelling strategy presents 24 explanatory variables to the regression modelling process – 10 sample estimates plus 14 complexity measures – from which different subsets will be expected to be used for each of the 10 learner models.

strategy	sys.mae	impr.perc	seconds	time.prop	efficiency
smp_d	0.042	95.0%	20,292	8.9%	9
smp_e	0.057	93.2%	32,092	14.1%	6
smp_c	0.107	87.2%	7,080	3.1%	24
smp_f	0.139	83.4%	4,046	1.8%	39
smp_a/complexity	0.198	76.4%	69,549	30.5%	2
smp_a/complex/landmarks	0.204	75.7%	69,729	30.6%	2
smp_a/all others	0.213	74.6%	70,049	30.7%	2
smp_a	0.226	73.1%	45,969	20.2%	3
smp_a/structural	0.226	73.0%	45,974	20.2%	3
smp_b	0.232	72.3%	5,879	2.6%	23
smp_a/complex lite/struct	0.235	72.0%	47,054	20.7%	3
all non-sample	0.328	60.8%	24,080	10.6%	5
complexity	0.339	59.6%	23,580	10.4%	5
complexity/structural	0.342	59.2%	23,585	10.4%	5
complexity lite	0.395	52.9%	1,080	0.5%	94
landmarks	0.436	48.0%	180	0.1%	510
statinfo	0.487	41.9%	40	0.0%	1,999
proposed	0.658	21.5%	250	0.1%	164
structural	0.805	4.0%	5	0.0%	1,541
tree	0.877	-4.7%	25	0.0%	0

Table 4.4: The relative performance of various accuracy modelling strategies

4.2.6 Modelling strategies at learner level

Table 4.5 presents a breakdown of the system MAEs of five selected modelling strategies by learner. Each cell in the first 5 columns represents the 10 x 10 CV MAE for that learner on the 50 datasets, for the modelling strategy indicated by the column header. The last column states the default MAE for that learner, which is the prediction error arising if the learners average classification error-rate is used as the prediction for all datasets.

learner	smp_d	smp_c	smp_a/cplx	smp_a	cplx	default.mae
oner	0.005	0.000	0.026	0.021	0.054	0.089
nbayes	0.002	0.000	0.024	0.022	0.061	0.085
j48	0.006	0.013	0.015	0.023	0.022	0.082
jrip	0.006	0.013	0.017	0.022	0.026	0.083
logistic	0.002	0.014	0.025	0.027	0.030	0.080
mlp100	0.005	0.018	0.017	0.026	0.020	0.079
svmpoly	0.002	0.024	0.017	0.021	0.025	0.084
bayesnet	0.004	0.000	0.020	0.021	0.043	0.085
knn	0.004	0.013	0.024	0.023	0.041	0.086
forest	0.005	0.014	0.013	0.019	0.018	0.086
system mae	0.042	0.107	0.198	0.226	0.339	0.838

Table 4.5: Mean absolute errors for five modelling strategies

4.2.7 Analysis of system MAE

MAE is an acceptable measure for comparing the accuracies of modelling strategies but the impact that a particular value will have on learner selection decisions is not obvious. The reason for this is that an average error tells us nothing about the error distribution. Table 4.6 provides an analysis for 3 selected modelling strategies with MAEs of approximately 4%, 10% and 20%, respectively. It shows that we might expect, for example, 90% of predictions to be within 1% of the actual error-rate with a system MAE of 4% but only 38% within 1% when the MAE is 20%.

	smp_d	smp_c	smp_a/cplx
mean system error	0.042	0.107	0.198
95% of predictions within	0.014	0.039	0.056
mean individual error	0.004	0.011	0.020
max individual error	0.034	0.146	0.293
predictions within 1%	90%	64%	38%
predictions within 2%	98%	80%	65%

Table 4.6: Mean error analysis for three modelling strategies

4.2.8 Sample estimates as predictors of the most accurate learner

Table 4.7 presents the results of an analysis undertaken to determine whether direct use of sample estimates to predict the most accurate learner was more reliable than using sample estimates in regression models.

The direct approach predicts as the most accurate learner the one with the highest sample accuracy. The model-based approach predicts an error-rate for each learner using the sample estimates and the learner with the lowest predicted error-rate is the prediction used.

The direct approach was tested with 25% samples and also with fixed samples of 200 instances. The regression models were trained with the smp_a set of variables, using a leave-one-out process, and the predictions for each dataset captured during the testing phase. The results are segmented to show the impact of the approaches on larger ($\geq$ 4,000 instances) and smaller datasets.

Two sets of regression models were produced; one set trained on all 50 datasets and another only trained on the 16 larger datasets. The model results for the smaller datasets are based on the 50-dataset models, whilst those for the larger datasets are based on the 16-dataset models. The p-values for tests of equal binomial proportions between the accuracies on the smaller and larger datasets for each approach are given, suggesting that only the accuracy of the 25% sample is influenced by dataset size.

dataset size	sample.25%	sample.200	reg.model
$\geq$4000 instances (16)	81.2%	25.0%	50.0%
<4000 instances (34)	23.5%	35.3%	23.5%
p-value	0.0004	0.6870	0.1219
all datasets (50)	42.0%	32.0%	32.0%

Table 4.7: The influence of dataset size on the proportion of datasets for which sample and model-based approaches predict the most accurate learner. The number of datasets are given in brackets. The p-value for a binomial test of equality between the proportions for each approach on large and small datasets is given.

Table 4.8 presents the regression model and 25% sample predictions for the larger datasets. The error columns state the difference between the error-rate of the learner that would have been selected using the relevant approach and the actual error-rate of the most accurate learner — what you would have lost by using the prediction. Aside from the first dataset (ID 100), the differences are sufficiently small to be deemed practically insignificant.

id	inst	actual	model	sample.25	model.error	sample.error
100	48,842	j48	jrip	j48	−0.016	0.000
114	19,020	forest	forest	forest	0.000	0.000
116	5,620	knn	mlp100	knn	−0.002	0.000
117	5,473	forest	forest	forest	0.000	0.000
118	10,992	mlp100	forest	mlp100	0.000	0.000
121	4,601	forest	forest	forest	0.000	0.000
125	5,000	mlp100	mlp100	mlp100	0.000	0.000
135	5,822	svmpoly	oner	knn	−0.001	0.000
136	4,521	logistic	forest	logistic	−0.008	0.000
137	43,500	forest	forest	forest	0.000	0.000
146	5,875	j48	j48	j48	0.000	0.000
147	7,128	mlp100	forest	logistic	−0.007	−0.004
149	20,640	forest	forest	forest	0.000	0.000
152	23,464	svmpoly	mlp100	svmpoly	−0.003	0.000
153	5,404	forest	forest	forest	0.000	0.000
156	8,993	j48	jrip	jrip	−0.002	−0.002

Table 4.8: Model vs 25% sample for predicting the most accurate learner on the larger datasets

4.3 Results time

4.3.1 Base-level data

Table 4.9 presents the details of 15 large datasets from the set used earlier. All datasets with at least 4,000 instances are included except for dataset ID 100 (UCI Adult) that was judged to have an evaluation time for svmpoly (135,358 s) that was an outlier with respect to the other dataset times. The times for a 25% sample estimate for all learners is given, along with the full evaluation time for all learners.

4.3.2 Non-sample variable group times

Table 4.10 presents the processing times for each group of non-sample explanatory variables, across the 15 large datasets. The time to obtain the full set of complexity measures was timed but the other times are estimates based on the pro-rata reduction in the total number of instances (approx. 31%) from the collection of 50 datasets. As the numbers are insignificant in the analysis this was felt to be adequate.

id	inst	attr	n1	maj	samp(s)	full (s)	full/samp
135	5,822	85	0.15	94.0	2,632	7,049	2.7
137	43,500	9	0.00	78.4	591	3,410	5.8
116	5,620	62	0.00	90.1	1,567	3,188	2.0
114	19,020	10	0.29	64.8	672	2,764	4.1
149	20,640	8	0.23	75.0	440	2,446	5.6
152	23,464	7	0.42	69.8	326	2,231	6.9
121	4,601	57	0.17	60.6	983	2,131	2.2
136	4,521	16	0.21	88.5	682	1,974	2.9
118	10,992	16	0.00	89.6	359	967	2.7
156	8,993	13	0.22	80.6	160	774	4.8
125	5,000	21	0.24	66.9	337	644	1.9
146	5,875	21	0.02	75.0	184	609	3.3
117	5,473	10	0.07	89.8	133	327	2.5
153	5,404	5	0.20	70.7	44	251	5.6
147	7,128	5	0.20	84.6	73	205	2.8
Total	n/a	n/a	n/a	n/a	9,183	28,970	n/a

Table 4.9: Details of the 15 larger diverse datasets with evaluation times

group	seconds
structural	3
statinfo	28
tree–based	17
landmarks	125
complexity	12,955
proposed	174
complexity lite	593
total	13,896

Table 4.10: Calculation times for non-sample explanatory variables on the 15 large datasets

4.3.3 Main result - modelling strategy performance

Table 4.11 presents the details of the accuracy versus time analysis for 5 selected time modelling strategies. The logarithm of the evaluation times was used as the variable to model in order to compensate for non-constant variance in the evalaution times (heteroscedasticity). Column 1 states the total MAE for the system of 10 learners on these datasets as determined by 10 x 10 CV processes. Column 2 states the MAE improvement as a percentage of the default MAE (8.8943). Columns 3 and 4 give the processing time for the strategy and the percentage of the full evaluation time that it represents. Column 5 gives the calculated efficiency measure defined earlier. The table has been ordered by system mae, column 1.

If we wanted to use regression models to predict the evaluation times of each

learner using 25% samples plus the structural measures, our modelling strategy would be smp_a/structural and using tables 4.9 and 4.10 we see that a processing time of 9,186 seconds or 31.7% of the full evaluation time would be required, delivering a mean error 63% better than using the average evaluation time for each learner as a constant prediction.

strategy	sys.mae	impr.perc	seconds	prop.time	efficiency
smp_a / structural	3.29	63.0%	9,186	31.7%	17.7
everything including smp_a	5.15	42.1%	23,080	79.7%	4.7
landmarks / structural	5.68	36.1%	128	0.4%	726.9
everything except samples	7.40	16.9%	13,897	48.0%	3.1
complexity / structural	8.63	2.9%	12,958	44.7%	0.6

Table 4.11: The relative performance of various time modelling strategies

4.3.4 Analysis of system MAE

Table 4.12 provides an analysis for 2 selected time modelling strategies. It shows that we might expect, for example, only 23% of predictions to be within 10% of the actual evaluation time if using the best strategy of smp_a / struct.

	smp_a/struct	land/struct
mean system error (log)	3.29	5.68
mean individual error (secs)	70	119
max individual error (secs)	1,825	3,296
95% of predictions within (secs)	310	660
within 1 minute of actual	80%	75%
within 2 minutes of actual	89%	84%
mean % error	33%	77%
within 10% of actual	23%	16%

Table 4.12: Mean error statistics for two time modelling strategies

4.3.5 Structure in evaluation times

50 homogenised datasets were prepared by randomly selecting 200 instances (100 per class) and selecting 3 continuous attributes from each of the datasets in the collection. The rationale was to remove the variation in time resulting from structural factors (instances, attributes etc), so only complexity remained as a variable. The evaluation times of these small datasets are too short and noisy for modelling. However, some structure emerged from an analysis of the time data that may prove useful for further research into modelling time.

Fig 4.1 shows the evaluation times for the *mlp100* learner on these datasets against n1, a key measure of boundary complexity. There is clearly 2 distinct

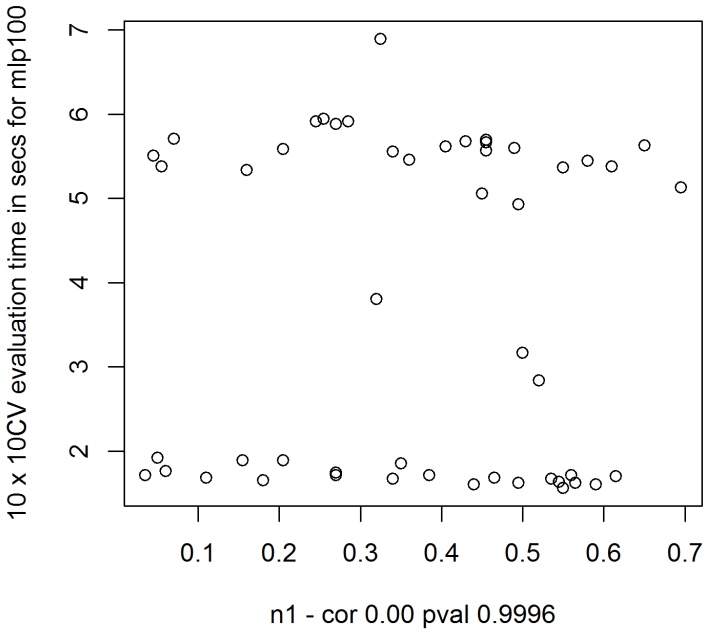

Figure 4.1: Structure in mlp100 evaluation time

groups of datasets, low and high evaluation times, but this group membership was not correlated with any of the available explanatory variables. A similar pattern was found for *svmpoly* but other learners did not show this structure. The *forest* learner is clearly correlated with complexity (Fig 4.2) but the others are not (e.g. Fig 4.3). These patterns were not detectable for the large datasets, where structural factors clearly have greater weight.

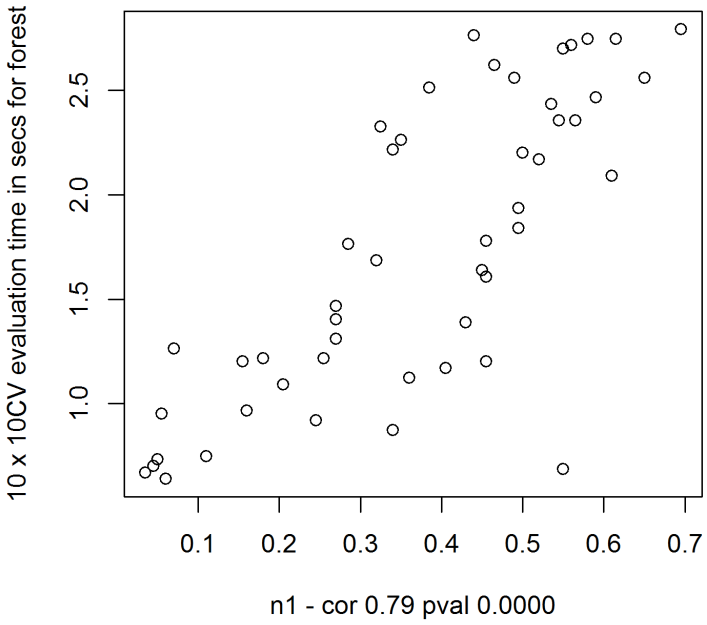

Figure 4.2: Structure in forest evaluation time

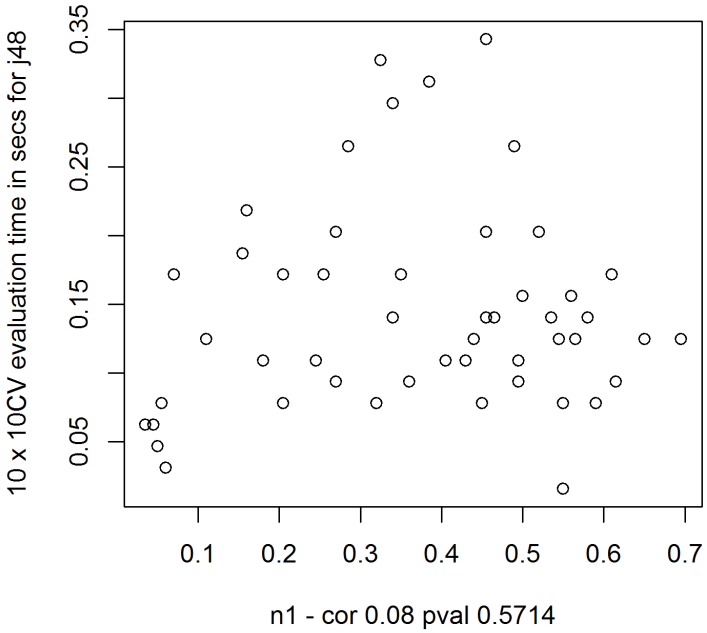

Figure 4.3: No structure in j48 evaluation time

4.4 Discussion

4.4.1 Accuracy

The performance of non-sample variables

It was hoped to determine whether regression models of learner accuracy need non-sample explanatory variables. The evidence presented here weighs against their continued use. Two contentions are advanced in support of a conclusion that they are not viable explanatory variables:

1. Modelling strategies based solely on non-sample variables cannot deliver a level of accuracy that is of practical use in learner selection.

2. Non-sample variables do not use potential evaluation time as efficiently as sample estimates

The first contention is based on the analysis of MAE in Section 4.2.7 and the results presented in Table 4.6. Thresholds of practical useability are clearly subjective and dependent upon operational considerations. In many fields though, even a 1% loss in classification accuracy could have important cost implications (e.g. fraud detection, medical diagnosis). So, models that have a mean error of more than 1%, alongside moderate likelihoods of errors much higher than this, are unlikely to be considered satisfactory in many real-world applications. Table 4.6 suggests that any modelling strategy that falls outside of a system MAE of 0.1 is unlikely to meet this practically useful criterion. The most accurate non-sample only strategy delivers a MAE of 0.3.

Another way in which to consider usefulness is in respect of discrimination ability - using the regression model predictions for deciding relative rankings. Fig 4.4 shows the distribution of the difference in error-rate between the 1st and 2nd learner by accuracy on each of the datasets. Around 75% of these differences are within 1% and 95% within 2%. Fig 4.5 shows the distribution of the differences in error-rate between the 1st and 8th learner (which generally excludes *oner* and *nbayes*, which are significantly less accurate than the other learners – see A3 for average rankings). Typically, only 5% separates the set of leading learners. Models with MAE above 1% (where only 64% of predictions are within 1%) will have little discriminatory power.

It is worth considering, even if non-sample only strategies are not viable, whether non-sample variables could make a useful contribution to predominately sample-based models. Complementing the smp_a strategy with the complexity measures produced a 16% relative improvement in MAE but at a cost of a 50% increase in relative processing time. The results in Table 4.7 illustrate the loss

of predictive power of sample estimates when the size falls to 200 or below. For smaller datasets, the predictive power of sample estimates relative to non-sample variables may be expected to decline. As many of the datasets in this experiment are small (below 4,000 instances) the impact that non-sample variables have, in the presence of sample estimates, is likely to be higher here than it will be in practical learner selection, where the datasets will be much larger.

The basis of the second contention is that it would appear that the most predictively powerful group of non-sample variables (the complexity measures) are simply not as time efficient as sample estimates. For example, smp_c delivers double the accuracy with just 10% of the time used by the smp_a/complexity strategy.

The case against using non-sample variables in learner accuracy prediction is strong.

Sample estimates and regression modelling

The question of whether regression modelling can increase the predictive power of sample estimates cannot be definitively answered here. There are three pieces of evidence to consider. Firstly, the results in Table 4.7 suggest that on larger datasets, direct prediction using sample estimates performs better than regression models built with the same variables. The difference in proportions is not significant (p=0.1365), there are only 16 data points, but it is large enough to be difficult to dismiss. However, the lack of training data for the regression models somewhat negates this apparent direct sample advantage.

Secondly, the sample error estimates on the 25% samples appear to be biased, with estimates tending to overstate the error on the full dataset. Fig 4.6 shows the bias evident in the *jrip* sample estimation, typical of many of the learners. This bias provides grounds for believing that there is non-random variation for models to explain. Hastie et al. (2011)[p. 243] describes how bias in CV estimates is a result of the interaction between sample size and the learning curve of a classifier – with large datasets, sample size would be more likely to be beyond the increasing phase of the learning curve and hence bias would be less significant and modelling less valuable. Results from an experiment with 50 large datasets would probably clarify the situation.

The final factor to consider is the ability of regression models to estimate learners that have not been sampled, using the sample estimates of other learners. Table 4.5 shows that the model for *mlp100* produced by smp_c is more accurate (0.018) than the model produced by smp_a (0.026), yet smp_c does not feature a sample of *mlp100*, saving considerable processing time. A single run of the modelling process on a smp_c variable set produced this model:

$$error\ rate\ mlp100 = 0.0015 + 0.5477 * forest.50 + 0.39 * logistic.50$$

This shows that the *mlp100* error rate can be successfully modelled (correlation 0.9692 / improvement on default MAE 77%) using estimates of the forest and logistic error-rates on 50% samples of the dataset. This is evidence of a synergy between the sample estimates of different learners that regression modelling has the potential to exploit.

There has been no tabulating of model structures nor analysis of the role of different variables on different learner models because the focus here has been on the modelling process. On another collection of datasets the 'best' model for *mlp100* using smp_b would certainly have different weights and would quite possibly use different sample estimates from the model stated above. McCullagh and Nelder (1989)[p. 8] warn that 'the data will often point with almost equal emphasis at several possible models'.

Inclusion of Dataset 100

The slow convergence of *svmpoly* on this dataset presented a dilemma - should it be excluded from the analysis given that the time for this one learner on one dataset was so significant? The decision was taken to leave it in because it may be the case that in every fifty or so datasets you encounter one that is slow to converge. Also, its inclusion was viewed as a bias in favour of the non-sample variables, which were not noticeably slower. The crude efficiency measure for smp_a falls from 3 to 1.3 and that for complexity from 4.8 to 2.7 when dataset 100 is removed. It's presence does not affect the conclusions drawn above.

4.4.2 Time

This was seen as an exploratory study and no firm aims were set. However, a number of issues became apparent that were not fully appreciated before hand. The most obvious of these, now, is the need for large(ish) datasets.The key issue with small datasets is that timing for fast learners is inaccurate. One of the reasons for only using the largest 15 datasets, was the suspicion that WEKA was understating the times for certain learners (*oner*, *nbayes*) on small datasets.

As highlighted in Chapter 2, few evaluation time regression studies have been published and the results presented here perhaps suggest why: it is difficult to build accurate models. This poor performance was not expected. At the outset it was intuitively felt that it could be the easier performance measure to model

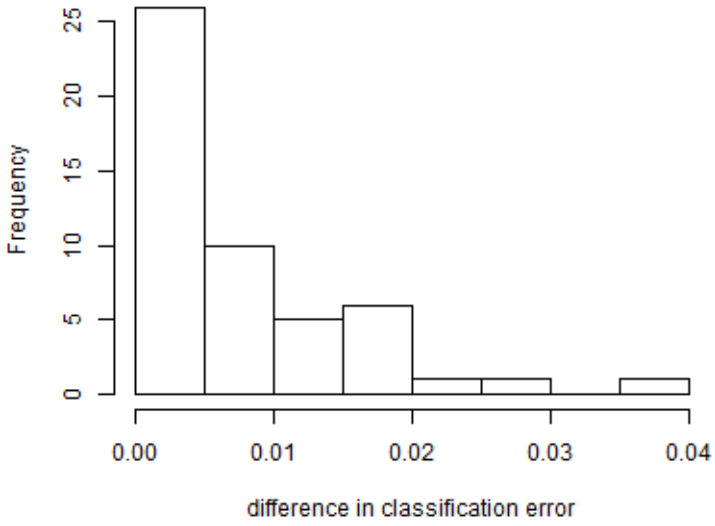

Figure 4.4: Mean absolute differences in error-rate between the 1st and 2nd most accurate learners

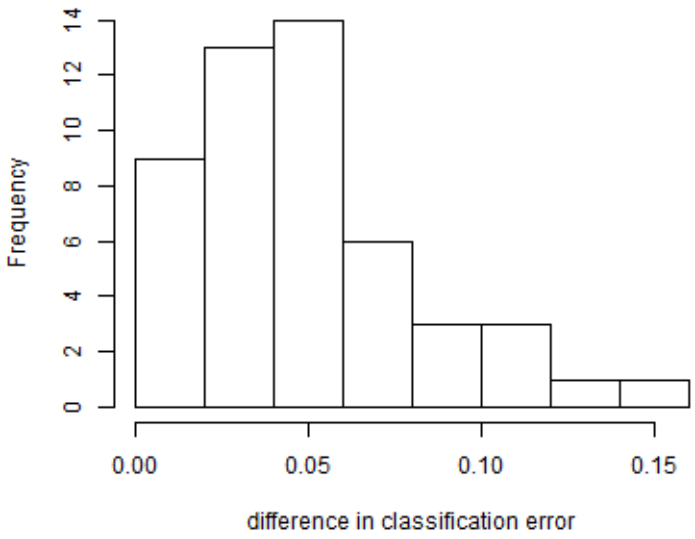

Figure 4.5: Mean absolute differences in error-rate between the 1st and 8th most accurate learners

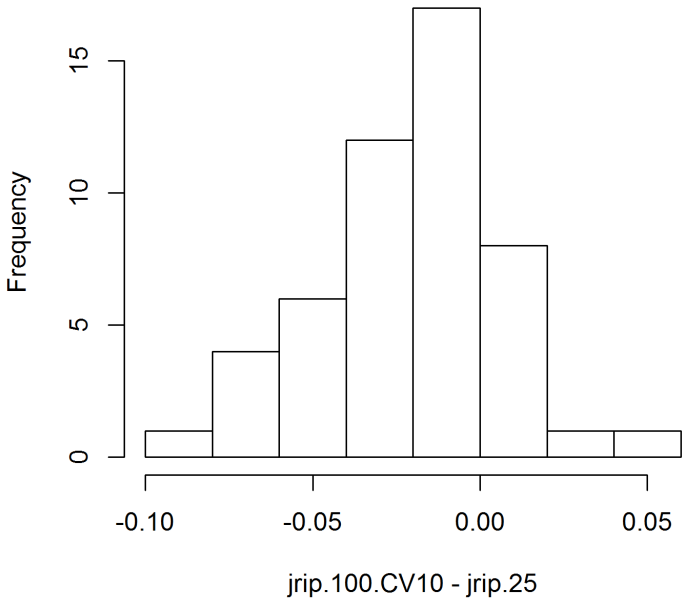

Figure 4.6: Bias in the 25 % jrip sample estimate

well; after all, relative evaluation times seem more stable than accuracies — *oner* is always faster than *mlp100*, for example.

Table 4.9 illustrates that there is no particular relationship between size and time nor between 25% sample time and that for full evaluation. Analysis of the large dataset times revealed no important relationships with any of the complexity variables. Table 4.12 shows that the best, sample-based, strategy produces predictions that are generally 33% in error. The modelling itself is hampered by the times becoming increasingly variable as they get larger, in addition to the order of magnitude differences in time between learners on the same dataset. Dataset 100 caused problems for *svmpoly* and to a lesser extent *mlp100* and this possibility of occasional, unusually slow convergence presents another challenge, although sample times may alert one to the problem.

Progress with predicting time accurately is likely to require a quite different approach to that taken for accuracy. The evidence of structure in the homogenised dataset results for *mlp100* may point the way forward. There may be some aspect of learner/data interaction that is not captured by external measures but that could be apparent at an early stage in the internal learner processes — perhaps the conditioning of the Hessian matrix or the state of other interim structures. Progress may require sophisticated learners to feedback performance data that is currently for internal consumption only.

4.4.3 General observations

The MAE analyses for accuracy and time need to be treated with some caution. They are based on an analysis of 500 predictions of a single realisation of the modelling process for the strategy being analysed. These predictions cannot be independent, as each dataset features 10 times. Also, it is not known whether the variance may be assumed to be constant from one strategy to another – in other words, 4% MAE for one strategy may have a different error distribution to a 4% MAE for another strategy – although there is no evidence of that effect here. However, there is clearly a practically important shift in reliability between the MAEs analysed.

Interpretation of the results would have been easier with a less crude measure of efficiency – increasing weight needs to be given to improvements in MAE as opposed to the current linear interpolation.

An important but difficult question is whether these results can be generalised beyond these 50 datasets. The answer is probably no. As discussed in the design choices, repository sets cannot be held to be representative of a wider set of problems — this is an issue that goes beyond this study. The only way to reach a safe conclusion is to perform several studies with different and larger collections of datasets and see if the pattern of results is consistently repeated.

4.5 Summary

The results demonstrate that non-sample explanatory variables such as statistical, information theoretic and complexity measures are redundant as explanatory variables in the regression modelling of learner accuracy.

The results also suggest that regression modelling could improve upon the efficiency of predicting accuracy using sample estimates only, although further experimental work is required.

It appears that producing reliable predictive models for evaluation time is more difficult than for accuracy and may require a different approach than regression modelling.

Chapter 5

Conclusions

5.1 The problem reconsidered

We began by placing a question mark over the viability of the type of meta-learning studies that have and continue to appear in the literature. Further discussion of the problem highlighted the lack of guidance available for data-mining professionals facing the problem of learner selection and its associated trade-offs. The need for a change in direction and emphasis was hinted at.

The weak link in these studies was identified as the indirect explanatory variable and in the Introduction the aim of establishing its efficacy was stated.

The evidence presented here firmly suggests that these indirect variables lack both the required predictive power and the time efficiency to contribute towards principled learner selection, which is the set of empirically established protocols that practitioners urgently need in this era of massive datasets.

The work here points towards basing the prediction of learner accuracy on sample-based models, although further work is required before firm conclusions can be reached.

5.2 Limitations of this work

Learner selection is a problem centred on big datasets. This study was conducted primarily on small datasets. This limitation does not undermine the core conclusions on variable efficacy but it has hampered efforts to establish whether direct samples were better than model-based solutions for predicting the most accurate learner.

5.3 Future work

Further research, on a larger collection of larger datasets, aimed at determining whether sample estimates used directly predict accuracy more reliably than modelled estimates, would seem a natural extension of the work here.

The evidence of this study suggests that accuracy can be predicted as precisely as required by selecting an appropriate sample strategy — the question for the practitioner is one of deciding how much model accuracy they are prepared to forego to save time. The big challenge to be overcome would appear to be that of reliably modelling learner evaluation time. The suggestion made here is that external variables, including sample-based estimates, do not appear to have the predictive power required to support precise learner selection decisions. There may be a limit to what the generalist can achieve here — experts in specific learning algorithms may be needed to open up the 'black-boxes' and provide early feedback to modelling systems from within the learning process itself.

There is also scope for research into the most appropriate decision-making frameworks for the learner selection problem. A starting point would be to explore just how efficient or otherwise ad-hoc processes are — how much accuracy is being lost? A benchmarking process for new approaches, based around the performance of ad-hoc selection, will be needed in the advance towards principled learner selection.

Appendix A

Data related to the diverse datasets

id	source	domain information
100	dcol - adl	salary from US census data
101	dcol - authors	author from words used
102	dcol - bal	scale inl balance
103	dcol - bpa	liver state
105	dcol - cmc	contraceptive method
106	dcol - col	lession type in horses
107	dcol - crx	outcome of credit card apps
108	dcol - drm	skin disease type
109	dcol - ecu	an aspect of eucalytus plants
110	dcol - h-s	presence of heart disease
111	dcol - ion	radar returns as good/bad
114	dcol - mag	signal type
116	dcol - opt	handwritten digits
117	dcol - pbc	page blocks from document image
118	dcol - pen	handwritten digits
119	dcol - pim	a persons diabetes class
120	dcol - seg	outdoor images
121	dcol - spa	spam emails
123	dcol - veh	vehicle silhouettes
125	dcol - wav21	waves
126	dcol - wbcd	cancer from mamograms
127	dcol - yea	protein site
128	uci - vertebral	orthopedic patients
129	uci - ilpd	liver disorder
130	uci - blood	blood donations
131	dcol - ann	steel annealing features - class changed to feature SHAPE
132	uci - mammographic	malignancy of mammographic masses
133	uci - steel	steel plate fault
134	uci - cardiotocography	fetal cardiotocograms
135	uci - insurance	caravan policy take-up
136	uci - bank	bank product take-up
137	uci - statlog	radiator positions in space shuttle
139	author	daily wholesale carpet sales as high or low
140	uci - housing	Boston house prices
141	uci - mpg	determine whether car is American
142	uci - auto	relative car prices
143	uci - computer	computer performance
144	uci - solar	historical solar flare complexity
145	uci - concrete	concrete compressive strength
146	uci - parkinsons	total UPDRS
147	liaad - ailerons	variance of f16 aileron positions
148	statlib - colleges	total sat scores
149	statlib - houses	house prices
150	statlib - irish	educational attainment
151	statlib - NO2	roadside NO2 levels
152	uofn - protein	protein structure co-ordination number - 10% sample 7 attrs
153	keel - phoneme	identification of nasal and oral sounds
154	keel - sa heart	heart disease in South African patients
155	keel - cylinder	identification of cylinder bands
156	keel - marketing	annual income of a household

Table A.1: Domain information

id	min	ave	max	range	stddev
100	0.139	0.161	0.190	0.051	0.0148
101	0.003	0.028	0.139	0.136	0.0417
102	0.043	0.122	0.366	0.323	0.0999
103	0.313	0.375	0.442	0.129	0.0540
105	0.296	0.332	0.368	0.072	0.0201
106	0.148	0.178	0.214	0.066	0.0213
107	0.136	0.154	0.220	0.084	0.0241
108	0.000	0.006	0.019	0.019	0.0062
109	0.051	0.084	0.128	0.077	0.0220
110	0.161	0.193	0.274	0.113	0.0350
111	0.072	0.123	0.179	0.107	0.0361
114	0.128	0.195	0.293	0.165	0.0566
116	0.000	0.009	0.058	0.058	0.0172
117	0.024	0.051	0.101	0.077	0.0229
118	0.001	0.017	0.065	0.064	0.0220
119	0.225	0.251	0.283	0.058	0.0173
120	0.003	0.023	0.162	0.159	0.0494
121	0.051	0.109	0.215	0.164	0.0563
123	0.178	0.240	0.307	0.129	0.0403
125	0.111	0.151	0.235	0.124	0.0338
126	0.025	0.055	0.119	0.094	0.0272
127	0.268	0.297	0.330	0.062	0.0212
128	0.149	0.194	0.237	0.088	0.0320
129	0.280	0.322	0.443	0.163	0.0459
130	0.216	0.236	0.273	0.057	0.0179
131	0.005	0.030	0.172	0.167	0.0517
132	0.173	0.191	0.217	0.044	0.0176
133	0.000	0.088	0.426	0.426	0.1663
134	0.043	0.087	0.143	0.100	0.0296
135	0.060	0.087	0.214	0.154	0.0523
136	0.099	0.112	0.128	0.029	0.0086
137	0.000	0.023	0.103	0.103	0.0334
139	0.163	0.202	0.236	0.073	0.0223
140	0.068	0.118	0.293	0.225	0.0651
141	0.084	0.139	0.269	0.185	0.0574
142	0.073	0.108	0.158	0.085	0.0287
143	0.041	0.095	0.169	0.128	0.0486
144	0.250	0.268	0.280	0.030	0.0089
145	0.058	0.141	0.234	0.176	0.0583
146	0.001	0.039	0.142	0.141	0.0441
147	0.104	0.117	0.143	0.039	0.0131
148	0.016	0.106	0.451	0.435	0.1279
149	0.093	0.129	0.170	0.077	0.0259
150	0.113	0.133	0.162	0.049	0.0149
151	0.200	0.226	0.251	0.051	0.0169
152	0.226	0.235	0.252	0.026	0.0084
153	0.097	0.190	0.253	0.156	0.0575
154	0.273	0.303	0.342	0.069	0.0239
155	0.284	0.351	0.430	0.146	0.0468
156	0.118	0.132	0.162	0.044	0.0135

Table A.2: Diverse dataset error statistics

id	oner	nbayes	j48	jrip	logistic	mlp100
100	0.190(10)	0.168(8)	0.139(1)	0.155(4)	0.149(2)	0.178(9)
101	0.139(10)	0.011(6)	0.040(8)	0.044(9)	0.010(5)	0.005(3)
102	0.366(10)	0.045(2)	0.156(8)	0.147(7)	0.048(4)	0.046(3)
103	0.438(9)	0.442(10)	0.341(4)	0.343(5)	0.313(1)	0.318(2)
105	0.343(7)	0.344(9)	0.309(2)	0.296(1)	0.325(3)	0.335(6)
106	0.186(6)	0.214(10)	0.148(1)	0.153(3)	0.191(9)	0.186(5)
107	0.145(4)	0.220(10)	0.145(3)	0.146(5)	0.147(6)	0.155(9)
108	0.019(10)	0.004(7)	0.010(8)	0.013(9)	0.003(4.5)	0.000(1.5)
109	0.076(4)	0.128(10)	0.066(2)	0.051(1)	0.082(6)	0.086(8)
110	0.274(10)	0.161(2)	0.220(9)	0.211(8)	0.163(3)	0.180(5)
111	0.179(10)	0.178(9)	0.106(5)	0.102(3)	0.123(7)	0.091(2)
114	0.293(10)	0.273(9)	0.149(3)	0.153(4)	0.209(7)	0.145(2)
116	0.058(10)	0.004(6)	0.006(7)	0.006(8)	0.007(9)	0.002(2)
117	0.054(7)	0.101(10)	0.027(2)	0.029(3)	0.051(6)	0.050(5)
118	0.065(10)	0.027(8)	0.003(5)	0.002(4)	0.015(7)	0.001(1)
119	0.283(10)	0.243(4)	0.255(7)	0.258(8)	0.225(1)	0.239(3)
120	0.027(9)	0.162(10)	0.007(6)	0.007(8)	0.003(1)	0.003(3)
121	0.215(10)	0.206(9)	0.073(3)	0.072(2)	0.074(4)	0.093(5)
123	0.251(8)	0.302(9)	0.234(6)	0.234(5)	0.208(2)	0.178(1)
125	0.235(10)	0.156(8)	0.170(9)	0.148(6)	0.144(5)	0.111(1)
126	0.119(10)	0.067(9)	0.066(8)	0.060(7)	0.053(5)	0.031(3)
127	0.300(6)	0.311(7)	0.278(3)	0.277(2)	0.321(9)	0.288(5)
128	0.237(10)	0.221(8)	0.185(4)	0.187(5)	0.149(1)	0.152(2)
129	0.332(8)	0.443(10)	0.321(7)	0.318(6)	0.280(1)	0.300(4)
130	0.240(7)	0.248(8)	0.218(2)	0.216(1)	0.228(5)	0.221(3)
131	0.006(3)	0.051(9)	0.007(5)	0.006(4)	0.017(7)	0.020(8)
132	0.182(5)	0.214(9)	0.178(4)	0.173(1)	0.174(3)	0.191(6)
133	0.375(9)	0.426(10)	0.000(1.5)	0.000(4)	0.001(5)	0.000(3)
134	0.143(10)	0.114(9)	0.060(3)	0.059(2)	0.089(6)	0.076(4)
135	0.061(3)	0.214(10)	0.061(4)	0.062(5)	0.063(6)	0.069(7)
136	0.116(8)	0.128(10)	0.106(3)	0.106(2)	0.099(1)	0.115(7)
137	0.052(9)	0.103(10)	0.000(2)	0.000(3)	0.029(7)	0.002(5)
139	0.222(8)	0.197(5)	0.180(2)	0.186(3)	0.196(4)	0.224(9)
140	0.114(7)	0.293(10)	0.089(4)	0.090(5)	0.091(6)	0.089(3)
141	0.123(5.5)	0.269(10)	0.088(2)	0.084(1)	0.122(4)	0.128(7)
142	0.116(8)	0.090(3)	0.101(6)	0.111(7)	0.155(9)	0.095(4)
143	0.165(9)	0.101(6)	0.041(1)	0.050(3)	0.082(5)	0.104(7)
144	0.276(8)	0.270(7)	0.250(1)	0.280(10)	0.262(3)	0.261(2)
145	0.234(10)	0.195(9)	0.073(2)	0.086(3)	0.179(7)	0.109(4)
146	0.051(7)	0.142(10)	0.001(1)	0.003(2)	0.066(9)	0.012(4)
147	0.143(10)	0.129(8)	0.109(4)	0.105(2)	0.108(3)	0.104(1)
148	0.043(3)	0.136(9)	0.051(4)	0.016(1)	0.017(2)	0.071(6)
149	0.169(9)	0.170(10)	0.102(2)	0.112(3)	0.123(5)	0.120(4)
150	0.142(8)	0.137(7)	0.113(1)	0.133(5.5)	0.122(3)	0.115(2)
151	0.251(10)	0.221(5)	0.239(8)	0.232(7)	0.213(3)	0.200(1)
152	0.244(9)	0.237(7)	0.239(8)	0.229(3)	0.226(2)	0.229(4)
153	0.253(10)	0.239(8)	0.135(2)	0.144(4)	0.250(9)	0.190(5)
154	0.342(10)	0.288(4)	0.301(6)	0.294(5)	0.273(1)	0.283(3)
155	0.340(4)	0.375(8)	0.306(3)	0.366(6)	0.371(7)	0.293(2)
156	0.123(3)	0.162(10)	0.118(1)	0.120(2)	0.130(6)	0.127(5)
Ave	0.187(8.01)	0.192(8.02)	0.132(4.07)	0.133(4.35)	0.140(4.73)	0.132(4.13)

Table A.3: Learner error and rankings by diverse dataset 1

id	svmpoly	bayesnet	knn	forest
100	0.150(3)	0.159(6)	0.166(7)	0.156(5)
101	0.006(4)	0.005(2)	0.003(1)	0.013(7)
102	0.077(5)	0.175(9)	0.043(1)	0.112(6)
103	0.420(7)	0.432(8)	0.388(6)	0.320(3)
105	0.331(5)	0.330(4)	0.368(10)	0.343(8)
106	0.173(4)	0.190(8)	0.188(7)	0.151(2)
107	0.151(7)	0.138(2)	0.136(1)	0.154(8)
108	0.000(1.5)	0.003(4.5)	0.002(3)	0.004(6)
109	0.077(5)	0.085(7)	0.112(9)	0.075(3)
110	0.161(1)	0.174(4)	0.187(6)	0.195(7)
111	0.119(6)	0.105(4)	0.151(8)	0.072(1)
114	0.209(6)	0.223(8)	0.163(5)	0.128(1)
116	0.002(4)	0.004(5)	0.000(1)	0.002(3)
117	0.062(8)	0.069(9)	0.045(4)	0.024(1)
118	0.013(6)	0.044(9)	0.001(2)	0.001(3)
119	0.232(2)	0.248(5)	0.271(9)	0.254(6)
120	0.004(4)	0.007(7)	0.006(5)	0.003(2)
121	0.095(6)	0.101(7)	0.108(8)	0.051(1)
123	0.251(7)	0.307(10)	0.228(4)	0.208(3)
125	0.142(4)	0.148(7)	0.123(2)	0.135(3)
126	0.025(1)	0.056(6)	0.030(2)	0.043(4)
127	0.312(8)	0.330(10)	0.281(4)	0.268(1)
128	0.212(7)	0.237(9)	0.195(6)	0.169(3)
129	0.286(2)	0.333(9)	0.310(5)	0.299(3)
130	0.238(6)	0.250(9)	0.225(4)	0.273(10)
131	0.012(6)	0.005(2)	0.172(10)	0.005(1)
132	0.206(8)	0.173(2)	0.203(7)	0.217(10)
133	0.000(1.5)	0.062(8)	0.006(6)	0.007(7)
134	0.099(7)	0.104(8)	0.083(5)	0.043(1)
135	0.060(1)	0.150(9)	0.060(2)	0.075(8)
136	0.107(5)	0.122(9)	0.113(6)	0.107(4)
137	0.030(8)	0.010(6)	0.002(4)	0.000(1)
139	0.203(6)	0.163(1)	0.211(7)	0.236(10)
140	0.086(2)	0.143(9)	0.122(8)	0.068(1)
141	0.123(5.5)	0.208(9)	0.143(8)	0.102(3)
142	0.098(5)	0.081(2)	0.158(10)	0.073(1)
143	0.135(8)	0.060(4)	0.169(10)	0.043(2)
144	0.269(6)	0.266(5)	0.266(4)	0.278(9)
145	0.184(8)	0.134(5)	0.156(6)	0.058(1)
146	0.064(8)	0.040(6)	0.013(5)	0.003(3)
147	0.114(7)	0.132(9)	0.111(6)	0.111(5)
148	0.063(5)	0.083(7)	0.451(10)	0.126(8)
149	0.133(7)	0.144(8)	0.127(6)	0.093(1)
150	0.128(4)	0.133(5.5)	0.148(9)	0.162(10)
151	0.250(9)	0.219(4)	0.224(6)	0.210(2)
152	0.226(1)	0.234(6)	0.233(5)	0.252(10)
153	0.227(6)	0.230(7)	0.136(3)	0.097(1)
154	0.277(2)	0.322(8)	0.316(7)	0.331(9)
155	0.348(5)	0.430(10)	0.399(9)	0.284(1)
156	0.130(7)	0.148(9)	0.132(8)	0.126(4)
Ave	0.146(5.15)	0.159(6.54)	0.158(5.74)	0.131(4.26)

Table A.4: Learner error and rankings by diverse dataset 2

id	boxM	sdr	lda	cancor	hotel	intercor	maxcor	avecor
100	0	4.47	0.20	0.47	9.56	0.06	0.33	0.19
101	0	5.98	0.00	0.95	8.98	0.15	0.70	0.26
102	1	1.01	0.05	0.79	6.96	0.00	0.40	0.40
103	0	1.16	0.30	0.37	3.97	0.26	0.16	0.09
105	0	1.04	0.37	0.22	4.34	0.54	0.12	0.11
106	0	1.13	0.33	0.30	3.61	0.12	0.24	0.11
107	0	4.01	0.26	0.48	5.33	0.17	0.41	0.23
108	0	9.99	0.00	0.97	8.53	0.23	0.91	0.39
109	0	7.72	0.10	0.73	6.70	0.13	0.49	0.20
110	0	1.47	0.15	0.74	5.77	0.16	0.53	0.31
111	0	9.99	0.10	0.79	6.34	0.23	0.52	0.18
114	0	1.08	0.22	0.57	9.13	0.28	0.46	0.16
116	0	1.31	0.00	0.88	9.84	0.12	0.57	0.15
117	0	1.32	0.05	0.67	8.38	0.29	0.42	0.17
118	0	1.15	0.03	0.73	9.43	0.27	0.58	0.28
119	0	1.16	0.22	0.55	5.81	0.17	0.47	0.21
120	0	1.92	0.02	0.77	8.14	0.28	0.47	0.16
121	0	1.39	0.11	0.75	8.67	0.06	0.38	0.16
123	0	5.91	0.19	0.54	5.86	0.41	0.26	0.15
125	0	1.35	0.14	0.66	8.26	0.30	0.56	0.22
126	0	9.99	0.04	0.88	7.57	0.39	0.79	0.47
127	0	1.46	0.32	0.37	5.47	0.09	0.28	0.11
128	0	9.99	0.14	0.58	5.07	0.41	0.44	0.33
129	0	9.99	0.28	0.34	4.35	0.21	0.25	0.16
130	0	1.13	0.23	0.36	4.71	0.47	0.28	0.19
131	0	8.47	0.15	0.66	6.55	0.08	0.62	0.23
132	0	1.60	0.19	0.65	6.53	0.22	0.56	0.37
133	0	2.17	0.26	0.47	6.29	0.25	0.32	0.11
134	0	2.03	0.10	0.74	7.85	0.24	0.49	0.17
135	1	1.29	0.06	0.27	6.13	0.07	0.15	0.04
136	0	1.11	0.11	0.42	6.89	0.07	0.40	0.11
137	0	1.04	0.09	0.78	11.09	0.19	0.67	0.32
139	0	1.01	0.25	0.15	3.35	0.33	0.12	0.09
140	0	9.99	0.08	0.73	6.33	0.39	0.62	0.32
141	0	9.99	0.12	0.73	6.11	0.56	0.72	0.38
142	0	9.99	0.08	0.80	5.89	0.38	0.75	0.45
143	0	9.99	0.08	0.78	5.74	0.51	0.64	0.47
144	0	9.99	0.37	0.19	3.69	0.20	0.17	0.12
145	0	1.16	0.18	0.61	6.44	0.21	0.40	0.20
146	0	1.29	0.07	0.74	8.89	0.41	0.72	0.07
147	0	1.05	0.11	0.57	8.14	0.19	0.51	0.24
148	0	3.17	0.07	0.73	7.32	0.29	0.68	0.31
149	0	1.08	0.13	0.63	9.52	0.31	0.55	0.13
150	0	1.01	0.30	0.23	3.31	0.21	0.18	0.18
151	0	1.14	0.21	0.48	5.00	0.16	0.33	0.13
152	0	1.02	0.23	0.51	9.03	0.15	0.42	0.26
153	0	1.43	0.24	0.50	7.48	0.13	0.33	0.26
154	0	1.18	0.27	0.45	4.78	0.22	0.37	0.21
155	0	1.79	0.34	0.37	4.47	0.08	0.21	0.07
156	0	1.16	0.13	0.60	8.52	0.16	0.47	0.23

Table A.5: Statinfo values 1-8 for the datasets

id	skew1	skew2	kurt1	kurt2	related	entropy	mutual	enattr	nsratio
100	4.44	1.90	25.00	8.23	1.00	2.42	0.07	11.39	33.76
101	0.79	1.19	1.14	2.71	0.89	3.47	0.13	7.59	26.53
102	0.46	0.53	0.97	0.86	1.00	2.32	0.12	8.03	17.72
103	1.38	1.74	3.41	5.68	0.33	2.86	0.05	18.00	51.51
105	0.64	0.76	1.02	1.76	0.89	1.84	0.03	33.60	61.86
106	0.89	1.12	3.97	3.80	0.73	2.09	0.07	12.88	27.30
107	3.47	3.04	19.32	22.76	0.87	1.74	0.09	10.95	18.26
108	4.00	2.48	11.53	16.04	0.88	1.28	0.19	4.69	5.74
109	2.57	0.75	25.00	2.03	0.95	2.85	0.18	4.43	14.74
110	1.10	0.81	1.95	1.16	0.77	1.94	0.11	9.09	16.78
111	0.28	0.96	0.80	0.70	1.00	3.38	0.25	3.77	12.54
114	0.87	0.70	3.07	2.05	1.00	5.69	0.08	12.10	72.64
116	3.85	5.58	13.64	25.00	0.81	2.49	0.06	8.23	43.20
117	5.29	5.80	25.00	25.00	1.00	2.47	0.06	7.52	38.10
118	1.25	0.53	2.46	1.01	0.94	5.68	0.11	4.36	50.41
119	1.23	0.89	3.67	2.52	1.00	3.49	0.09	10.75	39.22
120	1.24	3.15	7.96	25.00	0.89	3.65	0.17	3.38	19.88
121	12.17	12.20	25.00	25.00	0.96	1.06	0.07	14.30	14.72
123	1.18	0.35	5.23	0.87	0.89	4.01	0.07	11.24	54.48
125	0.03	0.24	0.31	0.26	0.90	5.44	0.07	13.41	78.62
126	1.35	1.32	5.83	4.41	0.93	3.36	0.32	2.98	9.51
127	5.03	3.20	25.00	24.03	0.62	2.91	0.04	22.43	72.00
128	1.01	0.62	6.46	1.09	1.00	3.10	0.15	6.01	19.52
129	2.91	2.41	21.86	12.99	0.80	2.36	0.05	16.46	43.96
130	1.99	2.07	9.02	5.50	0.75	3.13	0.06	13.81	53.56
131	3.05	3.22	4.21	23.22	0.67	0.83	0.07	14.02	10.63
132	1.23	5.01	4.83	25.00	1.00	2.01	0.15	6.76	12.61
133	3.39	2.33	25.00	24.45	0.88	2.81	0.06	14.84	43.74
134	2.01	3.39	9.88	25.00	1.00	3.34	0.10	7.79	33.15
135	7.15	4.27	25.00	25.00	0.60	1.39	0.00	102.26	435.52
136	3.50	2.27	25.00	11.40	0.88	2.13	0.02	24.56	100.51
137	16.62	4.60	25.00	25.00	0.78	2.86	0.22	3.40	11.95
139	0.05	0.21	1.10	1.26	0.89	2.10	0.05	16.00	40.38
140	1.55	1.90	5.15	7.78	0.92	2.91	0.17	4.73	16.15
141	0.98	1.63	1.58	2.22	0.86	3.08	0.21	4.84	13.92
142	0.93	0.79	2.08	1.28	0.88	2.33	0.30	3.38	6.89
143	1.64	2.53	4.71	8.32	1.00	2.12	0.35	2.84	5.01
144	8.13	6.97	25.00	25.00	1.00	0.94	0.05	18.14	16.55
145	0.91	0.74	2.49	1.62	1.00	3.72	0.11	7.32	32.51
146	2.67	3.43	17.81	25.00	0.95	4.18	0.09	8.59	43.28
147	0.28	0.37	0.95	1.60	0.80	5.10	0.10	6.03	48.49
148	1.50	1.55	8.15	5.77	0.94	3.41	0.16	4.95	19.83
149	2.24	2.03	21.17	12.92	1.00	5.11	0.07	11.53	71.65
150	0.12	0.31	0.30	0.21	0.80	2.15	0.17	5.12	11.44
151	0.42	0.73	1.63	2.13	0.71	3.84	0.07	10.85	50.44
152	0.30	1.00	0.26	1.44	1.00	3.36	0.08	11.49	42.70
153	0.92	0.71	1.47	0.95	1.00	5.24	0.16	5.41	31.49
154	1.07	1.01	2.85	1.74	0.67	3.22	0.07	14.24	48.22
155	2.02	1.56	16.36	9.24	0.84	2.73	0.06	15.32	41.65
156	1.08	1.34	2.42	2.19	1.00	1.86	0.08	8.54	21.33

Table A.6: Statinfo values 9-17 for the datasets

id	l.lda	l.knn	l.oner	l.nbay	l.stump	treeHW	treeNH	treeLW	treeHP
100	0.199	0.208	0.188	0.169	0.236	0.800	1.750	1.600	1.333
101	0.010	0.483	0.510	0.392	0.510	1.000	2.167	2.167	2.000
102	0.042	0.151	0.354	0.090	0.354	0.800	3.000	2.400	2.667
103	0.390	0.368	0.393	0.487	0.350	0.875	3.143	2.875	2.333
105	0.369	0.389	0.343	0.347	0.363	1.333	1.500	2.333	4.000
106	0.310	0.232	0.192	0.208	0.192	1.000	1.000	2.000	2.000
107	0.243	0.230	0.145	0.183	0.145	1.250	1.600	2.250	2.500
108	0.008	0.000	0.024	0.000	0.024	0.667	1.500	1.333	1.000
109	0.116	0.340	0.316	0.372	0.348	1.800	1.556	3.000	4.500
110	0.141	0.261	0.228	0.185	0.228	0.833	4.000	3.500	1.667
111	0.150	0.076	0.076	0.353	0.067	1.167	1.857	2.333	7.000
114	0.214	0.769	0.907	0.699	1.000	1.000	1.500	1.750	2.000
116	0.009	0.000	0.059	0.003	0.059	1.333	1.750	2.667	2.000
117	0.052	0.087	0.123	0.231	0.274	1.000	1.857	2.000	3.500
118	0.023	0.002	0.065	0.025	0.104	0.667	2.500	1.833	1.333
119	0.206	0.307	0.230	0.222	0.230	1.200	1.667	2.200	2.000
120	0.024	0.011	0.008	0.201	0.166	1.000	1.750	2.000	2.000
121	0.112	0.202	0.233	0.372	0.279	1.000	2.000	2.167	3.000
123	0.191	0.233	0.278	0.302	0.243	1.400	2.429	3.600	2.333
125	0.144	0.182	0.240	0.153	0.256	0.833	2.000	1.833	1.667
126	0.072	0.067	0.093	0.062	0.104	0.833	2.200	2.000	1.667
127	0.313	0.374	0.360	0.315	0.390	1.250	1.400	2.000	5.000
128	0.179	0.952	0.952	0.952	0.952	1.600	2.000	3.400	4.000
129	0.342	0.328	0.333	0.384	0.278	0.714	2.200	1.714	1.667
130	0.247	0.303	0.236	0.232	0.236	2.000	1.000	2.500	4.000
131	0.114	0.128	0.007	0.059	0.007	1.000	1.000	1.500	2.000
132	0.190	0.287	0.199	0.214	0.199	1.000	1.333	1.667	3.000
133	0.268	0.418	1.000	0.776	1.000	1.000	1.000	1.167	6.000
134	0.130	0.152	0.383	0.192	0.383	0.667	2.500	1.778	2.000
135	0.071	0.093	0.061	0.221	0.061	1.333	0.750	1.667	4.000
136	0.109	0.133	0.114	0.113	0.114	1.250	1.400	2.000	2.500
137	0.085	0.000	0.052	0.103	0.075	1.000	1.333	1.667	1.500
139	0.253	0.239	0.152	0.237	0.152	1.000	2.000	2.200	2.500
140	0.064	0.064	0.198	0.128	0.169	1.200	2.833	3.600	2.000
141	0.140	0.104	0.289	0.326	0.081	4.500	1.667	8.000	9.000
142	0.100	0.186	0.186	0.257	0.186	1.333	1.250	2.000	2.000
143	0.125	0.070	0.183	0.113	0.183	1.333	1.000	1.667	4.000
144	0.402	0.257	0.268	0.251	0.343	0.750	1.667	1.500	1.500
145	0.191	0.237	0.343	0.114	0.443	1.200	2.833	3.600	3.000
146	0.074	0.343	0.168	0.353	0.263	1.750	2.286	4.250	3.500
147	0.113	0.132	0.140	0.123	0.153	1.333	1.250	2.000	4.000
148	0.113	0.476	0.034	0.142	0.045	0.667	1.500	1.333	1.000
149	0.131	0.250	0.210	0.237	0.207	2.000	1.167	2.667	3.000
150	0.282	0.271	0.224	0.235	0.224	2.000	1.000	2.500	4.000
151	0.235	0.271	0.218	0.188	0.241	1.000	2.625	2.750	4.000
152	0.223	0.290	0.250	0.242	0.249	1.000	1.000	1.333	3.000
153	0.244	0.107	0.254	0.258	0.242	1.000	2.000	2.250	2.000
154	0.342	0.376	0.433	0.312	0.306	1.400	2.000	3.000	3.500
155	0.380	0.546	0.350	0.568	0.661	5.000	1.667	8.667	7.500
156	0.130	0.146	0.096	0.137	0.096	0.667	1.750	1.333	2.000

Table A.7: Landmark and tree-based values for the datasets

id	noise	overlap	outliers	clusters	clusprop	bayratio	lkratio	mnorm	minval
100	0.00	0.23	0.04	2	0.39	0.73	0.957	0	0.000
101	0.00	0.00	0.03	6	0.42	1.00	0.022	2	0.000
102	0.00	0.04	0.00	2	0.39	1.01	0.280	0	0.000
103	0.00	0.28	0.04	4	0.13	0.72	1.061	0	0.001
105	0.35	0.48	0.00	3	0.48	0.96	0.949	1	0.000
106	0.00	0.31	0.02	10	0.34	0.84	1.334	0	0.000
107	0.00	0.30	0.04	4	0.37	0.88	1.056	0	0.000
108	0.00	0.00	0.06	4	0.31	0.72		1	0.000
109	0.00	0.06	0.01	8	0.35	0.60	0.340	0	0.000
110	0.00	0.13	0.00	6	0.50	0.99	0.542	1	0.000
111	0.00	0.37	0.13	6	0.30	0.94	1.983	0	0.000
114	0.00	0.45	0.03	6	0.43	0.83	0.278	0	0.000
116	0.00	0.01	0.05	5	0.16	1.00		2	0.000
117	0.00	0.26	0.05	5	0.22	0.81	0.602	0	0.000
118	0.00	0.01	0.03	7	0.22	0.99	9.553	0	0.000
119	0.00	0.33	0.02	4	0.41	0.92	0.672	0	0.000
120	0.00	0.03	0.03	8	0.11	0.88	2.108	1	0.000
121	0.00	0.09	0.11	6	0.42	0.78	0.557	0	0.000
123	0.00	0.13	0.01	6	0.26	0.88	0.821	0	0.000
125	0.00	0.14	0.00	3	0.41	0.90	0.787	2	0.000
126	0.00	0.10	0.09	6	0.41	0.97	1.071	0	0.000
127	0.00	0.27	0.03	2	0.09	0.87	0.836	0	0.000
128	0.00	0.22	0.01	3	0.44	1.00	0.188	0	0.000
129	0.00	0.28	0.05	4	0.40	0.90	1.041	0	0.000
130	0.06	0.32	0.01	8	0.31	0.86	0.815	0	0.000
131	0.00	0.01	0.01	3	0.30	0.98	0.895	1	0.000
132	0.07	0.20	0.02	4	0.49	0.98	0.660	0	0.000
133	0.00	0.38	0.04	4	0.28	0.80	0.641	0	0.000
134	0.00	0.11	0.06	7	0.42	0.97	0.855	0	0.000
135	0.01	0.09	0.11	4	0.11	0.73	0.762	2	0.000
136	0.00	0.26	0.04	3	0.15	0.89	0.814	0	0.000
137	0.00	0.10	0.01	6	0.12	0.75	210.333	0	0.000
139	0.26	0.51	0.00	5	0.35	0.79	1.057	2	0.000
140	0.00	0.09	0.04	4	0.32	0.91	0.994	0	0.000
141	0.00	0.12	0.01	5	0.35	1.07	1.347	0	0.000
142	0.00	0.11	0.02	6	0.39	0.98	0.538	0	0.000
143	0.00	0.20	0.05	6	0.37	0.98	1.775	0	0.000
144	0.33	0.37	0.03	4	0.16	0.96	1.566	0	0.000
145	0.00	0.16	0.03	3	0.42	0.94	0.805	0	0.000
146	0.00	0.08	0.04	4	0.19	0.96	0.215	0	0.000
147	0.00	0.16	0.01	3	0.46	0.98	0.862	2	0.000
148	0.00	0.07	0.08	5	0.36	0.91	0.237	0	0.000
149	0.00	0.19	0.03	6	0.43	0.89	0.524	0	0.000
150	0.11	0.34	0.00	2	0.44	0.82	1.043	2	0.000
151	0.00	0.22	0.01	4	0.40	0.85	0.870	0	0.000
152	0.01	0.24	0.01	2	0.33	0.90	0.770	0	0.000
153	0.00	0.21	0.01	3	0.47	0.95	2.278	1	0.000
154	0.00	0.37	0.02	5	0.48	0.97	0.909	0	0.000
155	0.00	0.25	0.04	4	0.04	0.67	0.696	0	0.000
156	0.01	0.17	0.03	7	0.39	0.95	0.895	0	0.000

Table A.8: Proposed variable values for the datasets

id	oner.25	nbay.25	knn.25	forest.25	baynet.25	jrip.25	j48.25	log.25	mlp.25	svm.25
100	0.188	0.165	0.170	0.160	0.157	0.155	0.141	0.147	0.180	0.152
101	0.235	0.010	0.005	0.024	0.007	0.076	0.082	0.016	0.000	0.000
102	0.401	0.101	0.084	0.129	0.197	0.191	0.149	0.047	0.047	0.075
103	0.468	0.390	0.432	0.357	0.446	0.335	0.338	0.376	0.395	0.445
105	0.372	0.372	0.382	0.380	0.365	0.354	0.348	0.333	0.401	0.356
106	0.229	0.339	0.339	0.259	0.279	0.240	0.258	0.424	0.375	0.318
107	0.133	0.209	0.141	0.174	0.129	0.164	0.173	0.194	0.171	0.165
108	0.000	0.009	0.000	0.003	0.008	0.000	0.000	0.011	0.000	0.000
109	0.073	0.103	0.147	0.108	0.098	0.043	0.070	0.192	0.125	0.133
110	0.187	0.145	0.146	0.155	0.195	0.167	0.174	0.206	0.197	0.155
111	0.137	0.199	0.303	0.114	0.110	0.179	0.151	0.219	0.209	0.177
114	0.294	0.279	0.179	0.146	0.236	0.160	0.164	0.216	0.152	0.213
116	0.058	0.004	0.000	0.004	0.004	0.010	0.008	0.011	0.001	0.002
117	0.044	0.096	0.050	0.030	0.054	0.037	0.033	0.040	0.048	0.064
118	0.077	0.038	0.004	0.003	0.054	0.008	0.009	0.020	0.002	0.018
119	0.277	0.267	0.307	0.235	0.276	0.274	0.289	0.229	0.242	0.247
120	0.025	0.177	0.015	0.006	0.011	0.015	0.015	0.007	0.005	0.007
121	0.221	0.200	0.135	0.068	0.108	0.101	0.099	0.083	0.113	0.116
123	0.279	0.343	0.256	0.233	0.297	0.261	0.256	0.238	0.249	0.250
125	0.237	0.167	0.144	0.150	0.174	0.171	0.192	0.152	0.124	0.145
126	0.084	0.049	0.043	0.034	0.050	0.060	0.062	0.042	0.017	0.018
127	0.342	0.301	0.280	0.286	0.329	0.292	0.285	0.299	0.261	0.313
128	0.314	0.224	0.246	0.214	0.243	0.226	0.247	0.192	0.212	0.295
129	0.331	0.370	0.251	0.322	0.357	0.327	0.280	0.290	0.248	0.281
130	0.238	0.227	0.230	0.289	0.244	0.270	0.266	0.220	0.228	0.240
131	0.013	0.090	0.261	0.027	0.032	0.009	0.009	0.029	0.073	0.135
132	0.160	0.175	0.199	0.179	0.191	0.168	0.166	0.170	0.161	0.181
133	0.382	0.379	0.037	0.047	0.184	0.066	0.000	0.000	0.000	0.000
134	0.148	0.123	0.090	0.076	0.125	0.088	0.092	0.083	0.083	0.099
135	0.060	0.212	0.060	0.070	0.078	0.063	0.063	0.065	0.078	0.060
136	0.115	0.131	0.114	0.112	0.121	0.107	0.111	0.106	0.124	0.116
137	0.050	0.098	0.003	0.000	0.010	0.001	0.001	0.035	0.006	0.045
139	0.187	0.206	0.213	0.246	0.204	0.177	0.195	0.226	0.276	0.191
140	0.076	0.306	0.146	0.080	0.104	0.066	0.090	0.107	0.098	0.092
141	0.110	0.250	0.130	0.143	0.185	0.105	0.082	0.133	0.148	0.110
142	0.139	0.114	0.162	0.085	0.071	0.093	0.053	0.120	0.104	0.103
143	0.168	0.067	0.105	0.101	0.062	0.084	0.095	0.093	0.122	0.174
144	0.318	0.317	0.312	0.312	0.305	0.356	0.355	0.324	0.323	0.323
145	0.235	0.195	0.202	0.113	0.197	0.138	0.130	0.207	0.126	0.194
146	0.067	0.160	0.030	0.022	0.042	0.030	0.014	0.067	0.028	0.070
147	0.154	0.124	0.120	0.108	0.135	0.110	0.110	0.107	0.108	0.120
148	0.052	0.140	0.476	0.141	0.099	0.048	0.066	0.078	0.078	0.088
149	0.170	0.177	0.141	0.105	0.150	0.126	0.124	0.124	0.121	0.131
150	0.161	0.166	0.205	0.248	0.154	0.179	0.168	0.181	0.191	0.161
151	0.242	0.260	0.250	0.246	0.287	0.267	0.279	0.202	0.166	0.250
152	0.244	0.236	0.238	0.251	0.237	0.236	0.243	0.224	0.229	0.224
153	0.253	0.237	0.178	0.139	0.216	0.175	0.184	0.247	0.206	0.233
154	0.438	0.288	0.284	0.315	0.364	0.353	0.347	0.238	0.226	0.251
155	0.383	0.352	0.425	0.373	0.427	0.445	0.381	0.398	0.378	0.385
156	0.127	0.170	0.140	0.136	0.154	0.119	0.128	0.129	0.131	0.128

Table A.9: Error estimates based on 25% samples of the datasets

id	inst	actual	model	sample.25	actual.min	model.min	sample.min
100	48842	j48	j48	j48	0.139	0.122	0.141
101	841	knn	nbayes	mlp100	0.003	0.005	0.000
102	625	knn	knn	logistic	0.043	0.089	0.047
103	345	logistic	j48	jrip	0.313	0.288	0.335
105	1473	jrip	logistic	logistic	0.296	0.300	0.333
106	368	j48	jrip	oner	0.148	0.225	0.229
107	690	knn	bayesnet	bayesnet	0.136	0.120	0.129
108	366	mlp100	forest	oner	0.000	−0.005	0.000
109	736	jrip	jrip	jrip	0.051	0.044	0.043
110	270	svmpoly	forest	nbayes	0.161	0.133	0.145
111	351	forest	forest	bayesnet	0.072	0.097	0.110
114	19020	forest	forest	forest	0.128	0.126	0.146
116	5620	knn	forest	knn	0.000	−0.004	0.000
117	5473	forest	forest	forest	0.024	0.020	0.030
118	10992	mlp100	forest	mlp100	0.001	−0.005	0.002
119	768	logistic	forest	logistic	0.225	0.206	0.229
120	2310	logistic	forest	mlp100	0.003	−0.002	0.005
121	4601	forest	forest	forest	0.051	0.055	0.068
123	846	mlp100	forest	forest	0.178	0.206	0.233
125	5000	mlp100	forest	mlp100	0.111	0.130	0.124
126	569	svmpoly	forest	mlp100	0.025	0.023	0.017
127	1484	forest	mlp100	mlp100	0.268	0.240	0.261
128	310	logistic	forest	logistic	0.149	0.190	0.192
129	583	logistic	mlp100	mlp100	0.280	0.245	0.248
130	748	jrip	logistic	logistic	0.216	0.217	0.220
131	898	forest	jrip	j48	0.005	0.005	0.009
132	961	jrip	jrip	oner	0.173	0.145	0.160
133	1941	j48	knn	j48	0.000	0.035	0.000
134	2126	forest	forest	forest	0.043	0.063	0.076
135	5822	svmpoly	jrip	knn	0.060	0.052	0.060
136	4521	logistic	jrip	logistic	0.099	0.093	0.106
137	43500	forest	forest	forest	0.000	−0.008	0.000
139	1242	bayesnet	jrip	jrip	0.163	0.160	0.177
140	506	forest	jrip	jrip	0.068	0.062	0.066
141	398	jrip	jrip	j48	0.084	0.092	0.082
142	205	forest	j48	j48	0.073	0.044	0.053
143	209	j48	bayesnet	bayesnet	0.041	0.059	0.062
144	1066	j48	forest	bayesnet	0.250	0.278	0.305
145	1030	forest	forest	forest	0.058	0.097	0.113
146	5875	j48	forest	j48	0.001	0.014	0.014
147	7128	mlp100	forest	logistic	0.104	0.091	0.107
148	1302	jrip	j48	jrip	0.016	0.061	0.048
149	20640	forest	forest	forest	0.093	0.089	0.105
150	500	j48	bayesnet	bayesnet	0.113	0.143	0.154
151	500	mlp100	mlp100	mlp100	0.200	0.203	0.166
152	23464	svmpoly	logistic	svmpoly	0.226	0.203	0.224
153	5404	forest	forest	forest	0.097	0.121	0.139
154	462	logistic	mlp100	mlp100	0.273	0.254	0.226
155	539	forest	mlp100	nbayes	0.284	0.331	0.352
156	8993	j48	jrip	jrip	0.118	0.105	0.119

Table A.10: Model vs 25% sample for predicting the most accurate learner

Bibliography

S. Abdelmessih. Classifiers' accuracy prediction based on data characterization. Master's thesis, Multimedia Analysis and Data Mining Competence Center German Research Center for Artificial Intelligence (DFKI GmbH) Kaiserslautern, Germany, 2010.

D. Aha. Generalizing from case studies: A case study. In *In Proceedings of the Ninth International Conference on Machine Learning*, pages 1–10. Morgan Kaufmann, 1992.

J. Alcal-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sanchez, and F. Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17:255–287, 2011.

S. Ali and K. Smith. On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119–138, 2006.

A. Ben-David. About the relationship between roc curves and cohens kappa. *Engineering Applications of Artificial Intelligence*, 21:874–882, 2008.

H. Bensusan and C. Giraud-Carrier. Casa batlo is in passeig de gracia or landmarking the expertise space. In *Proceedings of the ECML'2000 workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination*, pages 29–47. ECML'2000, 2000.

A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

P. Brazdil and C. Soares. Zoomed ranking: Selection of classification algorithms based on relevant performance information. *Principles of Data Mining and Knowledge Discovery Lecture Notes in Computer Science*, 1910:160–181, January 2000.

P. Brazdil, R. Leite, J. Vanschoren, and F. Queiros. Using active testing and meta-level information for selection of classification algorithms. CW Reports

CW591, Department of Computer Science, K.U.Leuven, Aug 2010. URL https://lirias.kuleuven.be/handle/123456789/278649.

S. Cacoveanu, C. Vidrighin, and R. Potolea. Evolutional meta-learning framework for automatic classifier selection. In *Intelligent Computer Communication and Processing, 2009. ICCP 2009. IEEE 5th International Conference on*, pages 27–30, 2009.

M. Crawley. *Statistics: An introduction using R.* John Wiley & Sons,Ltd, 2005.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

J. Fürnkranz and J. Petrak. An evaluation of landmarking variants. In *Proceedings of the ECML/PKDD Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning (IDDM-2001)*, pages 57–68, 2001.

C. Giraud-Carrier and F. Provost. Toward a justification of metalearning: Is the no free lunch theorem a show-stopper? In *Proceedings of the ICML-2005 Workshop on Meta-learning*, pages 9–16, 2005.

S. Green. How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26:499–510, 1991.

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition, 2006.

B. Hanczar, J. Hua, C. Sima, J. Weinstein, and M. Bittner. Small-sample precision of roc-related estimates. *Bioinformatics*, 26(6):822–830, 2010.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2nd revised edition, 2011.

T. Ho and M Basu. Complexity measures of supervised classification problems. *IEEE Transactions on pattern analysis and machine intelligence*, 24:289–300, March 2002.

R. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–90, April 1993.

G. John and P. Langley. Static versus dynamic sampling for data mining. In *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 367–370. AAAI Press, 1996.

A. Kalousis. *Algorithm Selection via Meta-Learning*. PhD thesis, University of Geneve, Department of Computer Science, 2002.

A. Kalousis, J. Gama, and M. Hilario. On data and algorithms: Understanding inductive performance. *Machine Learning*, 54(3):275–312, March 2004.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference On Artificial Intelligence*, pages 1137–1143. Morgan Kaufmann, 1995.

C. Köpf, C. Taylor, and J. Keller. Meta-analysis: From data characterisation for meta-learning to meta-regression. In *Proceedings of the PKDD-00 Workshop on Data Mining, Decision Support,Meta-Learning and ILP*, 2000.

J. Lee and C. Giraud-Carrier. Predicting algorithm accuracy with a small set of effective meta-features. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 808–812. IEEE Computer Society, 2008. doi: 10.1109/ICMLA.2008.62.

R. Leite and P. Brazdil. Predicting relative performance of classifiers from samples. In *ICML '05 Proceedings of the 22nd international conference on Machine learning*, pages 497 – 503. ACM, 2005.

G. Lindner and R. Studer. Ast: Support for algorithm selection with a cbr approach. In *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, pages 418–423. Springer-Verlag, 1999.

H. Liu, F. Hussain, L. Chew, and D. Manoranjan. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6:393423, 2002.

J. Luengo and F. Herrera. Domains of competence of fuzzy rule based classification systems with data complexity measures: a case study using a fuzzy hybrid genetic based machine learning method. *Fuzzy Sets & Systems, Elsevier*, 161 (1):3–19, January 2010.

N. Macia. *Data Complexity In Supervised Learning A Far Reaching Implication*. PhD thesis, La Salle — Universitat Ramon Llull, 2011.

N. Macia, E. Bernado-Mansilla, A. Orriols-Puig, and T. Ho. Learner excellence biased by data set selection: A case for data characterisation and artificial data sets. *Pattern Recognition*, 2012.

E. Mansilla and T. Ho. Domain of competence of xcs classifier systemin complexity measurement space. *IEEE Transactions on Evolutionary Computation*, 1(9):82–104, February 2005.

K. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall, 2nd edition, 1989.

D. Michie, D. Spiegelhalter, and C. Taylor, editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

T. Mitchell. *Machine Learning*. McGraw-Hill, international edition, 1997.

A. Orriols-Puig, N. Macia, and T. Ho. Documentation for the data complexity library in c++. Technical report, La Salle - Universitat Ramon Llull, 2010.

Y. Peng, P. Flack, C. Soares, and P. Brazdil. Improved dataset characterisation for meta-learning. In *Proceedings of the 5th International Conference on Discovery Science*, pages 141–152. Springer-Verlag, 2002.

J. Petrak. Fast subsampling performance estimates for classification algorithm selection. In *Proceedings of the ECML-00 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination*, pages 3–14, 2000.

B. Pfahringer, H. Bensusan, and C. Giraud-Carrier. Landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML'2000*, pages 743–750. Morgan Kaufmann, June 2000.

F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *In Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453. Morgan Kaufmann, 1997.

F. Provost, D. Jenson, and T. Oates. Efficient progressive sampling. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 23–32. ACM, 1999.

M. Reif, F Shafait, and A Dengel. Prediction of classifier training time including parameter optimization. In *Proceedings of the 34th Annual German conference on Advances in artificial intelligence*, pages 260–271. Springer-Verlag, 2011.

M. Reif, F. Shafait, M Goldstein, T Breuel, and A Dengel. Automatic classifier selection for non-experts. *Pattern Analysis & Applications*, online:1–14, 2012.

J. Rice. The algorithm selection problem. *Advances in Computers*, 15:65–118, 1976.

S. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317328, 1997.

C Schaffer. A conservation law for generalization performance. In *In Proceedings of the 1994 International Conference on Machine Learning*, pages 259–265. Morgan Kaufmann, 1994.

S. Singh. Multiresolution estimates of classification complexity. *IEEE Transactions on pattern analysis and machine intelligence*, 25:1534–1539, December 2003.

J. Smith, M. Tahir, D. Sannen, and H. van Brussel. Making early predictions of the accuracy of machine learning applications. *ArXiv e-prints*, 2012. URL `http://adsabs.harvard.edu/abs/2012arXiv1212.1100S`.

K. Smith-Miles. Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys*, 41(1):6:1–6:25, Dec 2008.

S. Sohn. Meta analysis of classification algorithms for pattern recognition. *IEEE Transactions on pattern analysis and machine intelligence*, 21(11):1137–1144, 1999.

L. Trujillo, E. Galvan-Lopez, and P. Legrand. Predicting problem difficulty for genetic programming applied to data classification. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, GECCO'11, pages 1355–1362, Dublin, Ireland, July 2011. ACM.

K. Tsiptsis and A. Chrorianopoulos. *Data Mining Techniques in CRM*. John Wiley & Sons,Ltd, 2009.

I. Witten, E. Frank, and M. Hall. *Data Mining : practical machine learning tools and techniques*. Morgan Kaufmann, 3rd edition, 2011.

D. Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6:47–94, 1992.

Y. Yang and G. Webb. Proportional k-interval discretization for naive-bayes classifiers. In *Proceedings of the 12th European Conference on Machine Learning*, pages 564–575. Springer-Verlag, 2001.